Improved data analysis for the MinION nanopore sequencer

Miten Jain[1,2], Ian T Fiddes[1,2], Karen H Miga[1,2], Hugh E Olsen[1,2], Benedict Paten[1,2] & Mark Akeson[1,2]

Speed, single-base sensitivity and long read lengths make nanopores a promising technology for high-throughput sequencing. We evaluated and optimized the performance of the MinION nanopore sequencer using M13 genomic DNA and used expectation maximization to obtain robust maximum-likelihood estimates for insertion, deletion and substitution error rates (4.9%, 7.8% and 5.1%, respectively). Over 99% of high-quality 2D MinION reads mapped to the reference at a mean identity of 85%. We present a single-nucleotide-variant detection tool that uses maximum-likelihood parameter estimates and marginalization over many possible read alignments to achieve precision and recall of up to 99%. By pairing our high-confidence alignment strategy with long MinION reads, we resolved the copy number for a cancer-testis gene family (CT47) within an unresolved region of human chromosome Xq24.

© 2015 Nature America, Inc. All rights reserved.

In 2014, Oxford Nanopore Technologies (ONT) enlisted several hundred laboratories to beta-test its 100-gram MinION sequencing device. The MinION sequences individual DNA molecules, providing very long read lengths to help overcome some of the drawbacks of short-read sequencing. As part of the MinION Access Program (MAP), we set out to characterize the performance and characteristics of the sequencing platform and to develop it to call single-nucleotide variants (SNVs) and resolve the repeat structure of highly repetitive regions. Our open-source analysis tools are available online (**Supplementary Software 1** and **2**; https://github.com/mitenjain/nanopore and https://github.com/benedictpaten/marginAlign for the nanopore and marginAlign pipelines, respectively).

The MinION reads the sequences of individual DNA strands as they are driven through biological nanopores by an applied electric field. The rate at which each DNA strand moves through a nanopore is controlled by a processive enzyme bound to the DNA at the pore orifice. Up to 512 DNA molecules can be read simultaneously using amplifiers that independently address each nanopore. Changes in ionic current, each associated with a unique five-nucleotide DNA *k*-mer, are detected as DNA molecules translocate through the nanopores at single-nucleotide precision. DNA bases are called using cloud-based software (Metrichor) provided by ONT that employs hidden Markov models (HMMs) to infer sequences from these current changes.

We determined MinION sequence-read quality and errors by analyzing the genome of M13mp18, a phage from *Escherichia coli* host strain ER2738 with a 42% average GC content and a 7.2-kb genome (Online Methods). Using expectation maximization, we inferred maximum-likelihood estimates (MLEs) for the rates of insertions, deletions and substitutions in MinION reads. We then realigned the reads to generate high-confidence alignments and used the MLE models to demonstrate that MinION reads can be used for accurate SNV calling. By coupling this alignment strategy with long MinION reads, we resolved the tandem-repeat organization of a CT47 cancer-testis gene family on an unfinished segment of human chromosome Xq24. Our results document the substantial improvements in the MinION's performance achieved during MAP.

RESULTS

The MinION reads both strands of duplex DNA

We prepared libraries as recommended by ONT, with modifications to ensure the integrity of high–molecular weight DNA (Online Methods). A DNA construct analyzed on the MinION (**Fig. 1**) is composed of a lead adaptor that loads the processive enzyme and facilitates DNA capture in the applied electric field; the DNA insert of interest; a hairpin adaptor that permits consecutive reading of the template and complement strands by the nanopore; and a tethering adaptor that concentrates DNA at the membrane surface.

Translocation of a single M13 genomic double-stranded DNA (dsDNA) copy through a MinION pore involves a series of steps, each associated with an identifiable ionic current pattern (**Fig. 1**). These include (i) the open pore; (ii,iii) capture and translocation of the lead adaptor; (iv) translocation of the template strand; (v) translocation of the hairpin adaptor; (vi) translocation of the complement strand (giving two-directional or 2D sequence data); (vii) translocation of the tethering adaptor; and (viii) release of the DNA strand into the *trans* compartment and return to the open-channel ionic current. At this point another DNA molecule can be captured and analyzed by the pore.

Over the first 6-month period of MAP, three MinION chemistry versions and numerous base-calling algorithm updates resulted in successive improvements in device performance (**Supplementary Fig. 1**). The average observed identity (the proportion of bases in a read that align to a matching base in a reference sequence) for

[1]UC Santa Cruz Genomics Institute, Santa Cruz, California, USA. [2]Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA. Correspondence should be addressed to B.P. (benedict@soe.ucsc.edu) or M.A. (makeson@soe.ucsc.edu).

RECEIVED 12 DECEMBER 2014; ACCEPTED 20 JANUARY 2015; PUBLISHED ONLINE 16 FEBRUARY 2015; DOI:10.1038/NMETH.3290

QIAGEN Exhibit 1028.0001

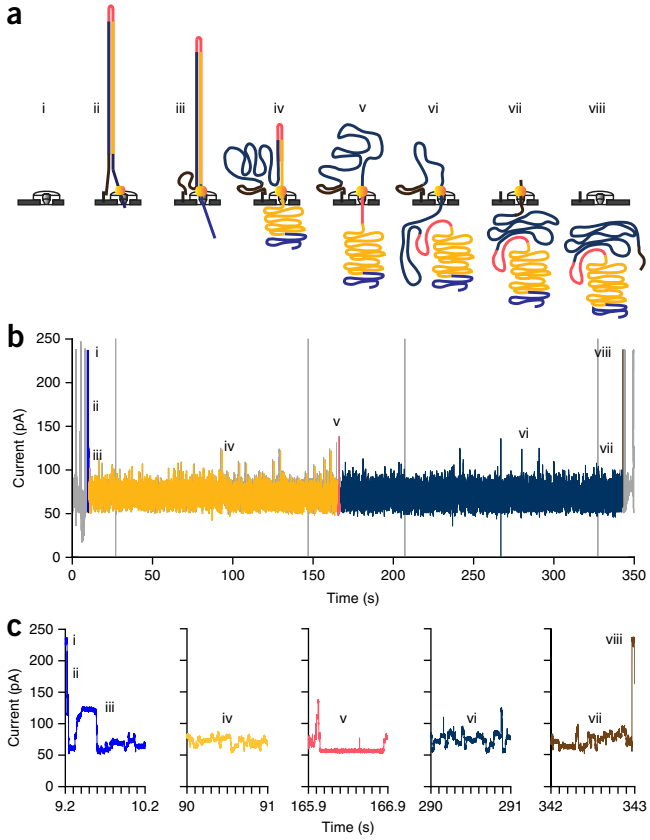


Figure 1 | Molecular events and ionic-current trace for a 2D read of an M13 phage dsDNA molecule. (**a**) Steps in DNA translocation through the nanopore: (i) open channel; (ii) dsDNA with lead adaptor (blue), bound molecular motor (orange) and hairpin adaptor (red) is captured by the nanopore; capture is followed by translocation of the (iii) lead adaptor, (iv) template strand (gold), (v) hairpin adaptor, (vi) complement strand (dark blue) and (vii) trailing adaptor (brown); and (viii) status returns to open channel. (**b**) Raw current trace for the passage of the M13 dsDNA construct through the nanopore. Regions of the trace corresponding to steps i–viii are labeled. (**c**) Expanded time and current scale for raw current traces corresponding to steps i–viii. Each adaptor generates a unique current signal used to aid base calling.

2D reads was 66% in June 2014 (R6.0 chemistry release), 70% in July 2014 (R7.0 chemistry release), 78% in October 2014 (R7.3 chemistry release) and 85% in November 2014 (Metrichor R.7X 2D version 1.9 update). The present study was based on MinION R7.3 chemistry and R7.X version 1.9 base-calling algorithms.

MinION throughput

We sequenced intact replicative-form M13 phage dsDNA using three MinION flow cells that contained 337–473 functional channels (Online Methods). Reads were characterized as 'template', 'complement' or '2D', with '2D' representing reads obtained by computationally merging template and complement data from the same hairpin-linked molecule. Each 48-h replicate run generated between 184 million and 450 million bases from 63% template, 24% complement and 13% 2D reads (**Supplementary Table 1**). Results presented in this paper are based on reads classified by Metrichor as high quality, which totaled between 60 million and 189 million bases per M13 sequencing run.

Establishing a mapping pipeline for MinION reads

To evaluate the quality of these reads, we experimented with four different alignment programs[1–4] (Online Methods). Each was run with its default parameters and with tuned parameters that were selected on the basis of experimentation or expert advice from other participants in MAP (**Supplementary Table 2**).

The proportion of reads that mapped to reference sequences (M13 or ONT λ DNA control) varied by aligner (**Supplementary Fig. 2**). LAST[3] with tuned parameters was the most inclusive program, and stringency analysis indicated that few of its alignments were false positives (**Supplementary Fig. 3**). For data pooled from the three M13 experiments, tuned LAST mapped 95.3% of template, 98.3% of complement and 89.9% of 2D reads. Most unmapped reads were homologous to *E. coli*, indicating minor contamination[5,6] (Online Methods, **Fig. 2a–c** and **Supplementary Table 3**).

We observed distinct peaks at 7.2 kb, corresponding to full-length M13 DNA, and at 3.8 kb, corresponding to the ONT λ phage DNA control (**Fig. 2a–c**). A large number of reads spanned the full M13 genome, whereas unmappable reads made up a small proportion (<0.2% of all 2D reads) and were generally shorter than mappable reads.

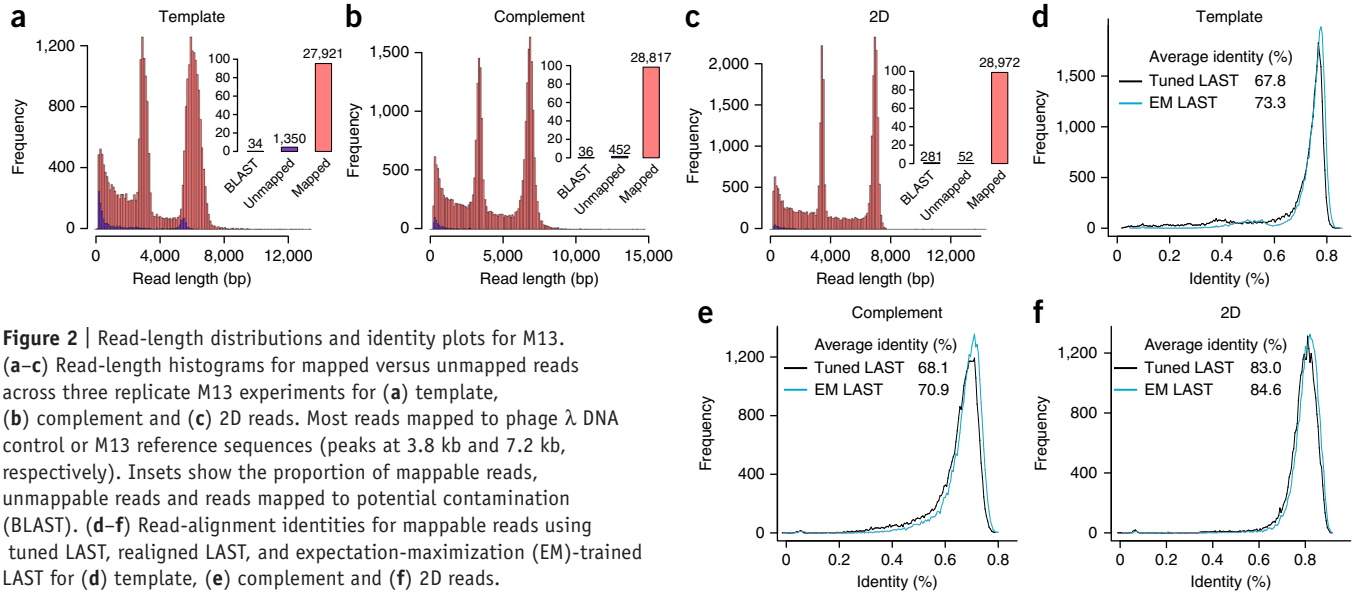


Figure 2 | Read-length distributions and identity plots for M13. (**a–c**) Read-length histograms for mapped versus unmapped reads across three replicate M13 experiments for (**a**) template, (**b**) complement and (**c**) 2D reads. Most reads mapped to phage λ DNA control or M13 reference sequences (peaks at 3.8 kb and 7.2 kb, respectively). Insets show the proportion of mappable reads, unmappable reads and reads mapped to potential contamination (BLAST). (**d–f**) Read-alignment identities for mappable reads using tuned LAST, realigned LAST, and expectation-maximization (EM)-trained LAST for (**d**) template, (**e**) complement and (**f**) 2D reads.

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0002

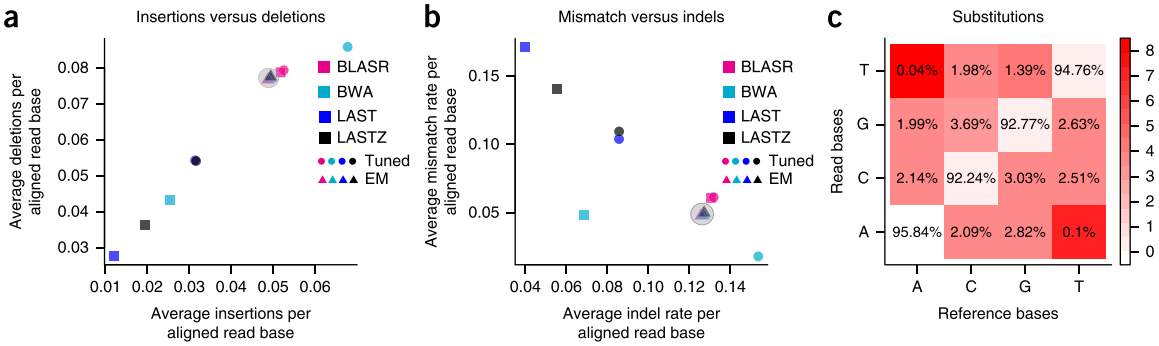


Figure 3 | Maximum-likelihood alignment parameters derived using expectation maximization (EM). The process starts with four guide alignments, each generated with a different mapper using tuned parameters. (**a**) Insertion versus deletion rates, expressed as events per aligned base. (**b**) Indel events per aligned base versus rate of mismatch per aligned base (Online Methods). Rates varied strongly between different guide alignments; however, EM training and realignment resulted in very similar rates (gray shading in circles), regardless of the initial guide alignment. (**c**) The matrix for substitution emissions determined using EM reveals very low rates of A-to-T and T-to-A substitutions. The color scheme is fitted on a log scale, and the substitution values are on an absolute scale.

© 2015 Nature America, Inc. All rights reserved.

Expectation maximization generates high-confidence read alignments

We found substantial disagreement among rates of substitution, insertion and deletion for alignments generated by different mapping programs (**Fig. 3a,b**). A more principled way to estimate true error rates is to propose a reasonable model of the error process and calculate MLEs of the parameters (Online Methods)[7]. Using expectation maximization to train an HMM (**Supplementary Fig. 4**) and alignment-banding heuristics for efficiency[8], we obtained robust convergence of parameter MLEs across all replicate experiments, guide alignments and random starting parameterizations (**Fig. 3a,b** and **Supplementary Fig. 5**). This showed that insertions were less frequent than deletions by about twofold in 2D reads and about threefold in template and complement reads. The combined insertion-deletion (indel) rate was between 0.13 (2D reads) and 0.2 (template and complement reads) events per aligned base. For all read types, indels were predominantly single bases (**Supplementary Fig. 6**). Substitutions varied from 0.21 (for template reads) to 0.05 (for 2D reads) events per aligned base (**Fig. 3c** and **Supplementary Figs. 7** and **8**). Substitution errors were not uniform; in particular, A-to-T and T-to-A errors were estimated to be very low, at 0.04% and 0.1%, respectively (**Supplementary Note 1**).

Realigning reads using the MLE parameters and the AMAP objective function[9] yielded substantial improvements over the initial alignments for every tuned program (Online Methods, **Fig. 2d–f** and **Supplementary Fig. 9**). For high-confidence alignments, there were no clear correlations between read length and errors (**Supplementary Fig. 10**). However, there were positive correlations among the rates of insertions, deletions and substitutions in 2D reads (**Supplementary Fig. 11** and **Supplementary Note 2**).

We also analyzed our data using a newly available Burrows-Wheeler Aligner (BWA) mode (ont2d) optimized for nanopore reads. The average percent identity obtained with ont2d was slightly less than the value obtained through expectation maximization (**Supplementary Table 4**); however, error rates were substantially closer to the MLE parameters estimated by expectation maximization, which suggests that ont2d is an improvement over the pacbio mode (for Pacific Biosciences) that we used originally.

To see whether our analysis pipeline produced similar results with larger, more complex genomes, we analyzed the *E. coli* data set released by Quick *et al.*[10], which used R7.3 chemistry and Metrichor R7.3 2D version 1.5. The most recent Metrichor update was not available when Quick *et al.*[10] released their data set. We observed an improvement in average identity from 80.1% with tuned LAST to 81.8% after realignment using the AMAP objective function with MLE parameters. In addition, the MLEs for the rates of insertions (0.0598 events per aligned base), deletions (0.0910) and substitutions (0.0531) were very similar to those found for the M13 data.

M13 sequencing depth and *k*-mer analysis

Sequencing depth was generally consistent across the 7.2-kb M13 genome (**Fig. 4** and **Supplementary Fig. 12**); however, 192 positions (2.6%) were underrepresented (**Supplementary Note 3**). Approximately 50% of these positions appeared at the beginning and end of the reference, and were likely the result of adaptor trimming by Metrichor. A majority of the remaining underrepresented positions were associated

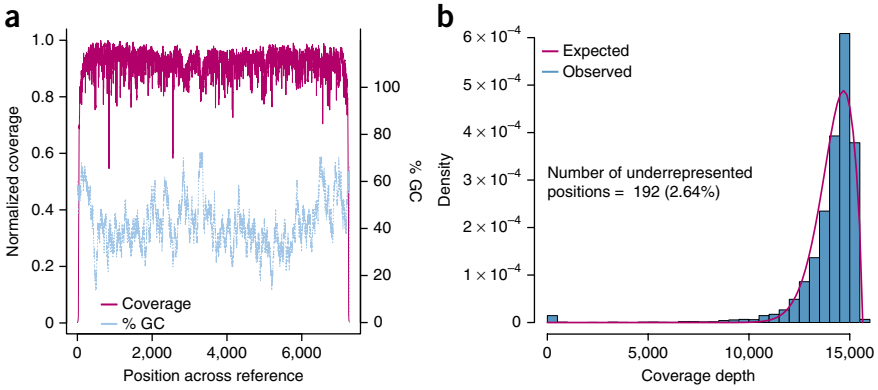


Figure 4 | M13 sequencing depth. (**a**) The magenta line denotes coverage by position in the genome (binned over a sliding 5-bp window), and the blue line depicts the local percentage of GC for that position (binned over a 50-bp sliding window). (**b**) Coverage-depth distribution fitted with a generalized extreme-value distribution.

QIAGEN Exhibit 1028.0003

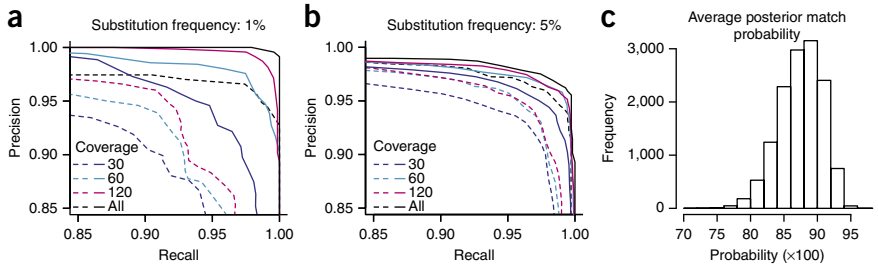


Figure 5 | Exploring SNV calling with MinION reads. (**a**,**b**) Variant calling with substitution frequencies of (**a**) 1% and (**b**) 5%. Dashed lines in both **a** and **b** represent results from variant calling using a transducer model conditioned on a fixed, tuned LAST alignment. Different sampled read coverages are shown. Each curve was produced by varying the posterior base-calling threshold to trade precision for recall. Solid lines in both **a** and **b** represent results from variant calling using the same transducer model as used for the tuned LAST alignments but incorporating marginalization over the read to reference alignments using a trained alignment model. Results shown are averaged over three replicate M13 experiments and, for each coverage level, three samplings of the reads. The "All" curve reflects all the available data for each experiment. (**c**) The distribution of posterior match probabilities shows that there was substantial uncertainty in most matches and demonstrates that marginalizing over the read alignments is a powerful approach.

with 5-mers rich in polymeric nucleotide runs (**Supplementary Table 5**). To determine whether the MinION has an inherent bias toward certain *k*-mers, we compared counts of 5-mers for all three read types (template, complement and 2D) with the M13 reference sequence. The most underrepresented 5-mers were homopolymers of poly(dA) or poly(dT), whereas the most overrepresented 5-mers were GC-rich and absent homopolymer repeats (**Supplementary Note 3** and **Supplementary Table 6**). These findings are consistent with observations from Ashton *et al.*[11].

MinION reads can call SNVs with high recall and precision

SNV detection is important for metagenomics and microbial-strain detection[12–14]. To determine whether MinION reads could be used for SNV discovery in monoploid genomes, we computationally introduced random substitutions into the M13 reference sequence at 1%–20% frequency. Using this altered sequence as an alignment reference, we attempted to recover these substitutions using a Bayesian transducer framework[15] (Online Methods and **Supplementary Note 4**) and assessed performance in terms of precision, recall and *F*-score. These experiments also addressed the accuracy of our alignments and error models while avoiding issues of reference-allele bias, to which simple metrics, such as alignment identity, are prone.

Using all the 2D read data and a posterior base-calling threshold that gave the optimal *F*-score, we achieved a recall of 99% and precision of 99% at 1% substitution frequency (**Fig. 5a**). When we reduced the sequencing depth down to a more reasonable 60× by sampling, we achieved recall and precision of 97%. Increasing the mutation frequency decreases the *F*-score progressively, presumably because alignment between the reads and the mutated reference becomes more difficult (**Fig. 5b**).

One particularly powerful strategy that we employed was marginalization over many possible alignments for each read, which helped factor out the considerable alignment uncertainty (**Fig. 5c**). In contrast, using fixed LAST alignments but otherwise keeping the method the same resulted in substantially higher rates of false positives for a given recall value (**Fig. 5a**,**b**).

Resolving the organization of a cancer-testis gene family

A strength of the MinION device is its ability to produce long, single-molecule reads. In addition to routinely observing full-length 2D reads of M13 genomic DNA (**Fig. 2**), we found substantially

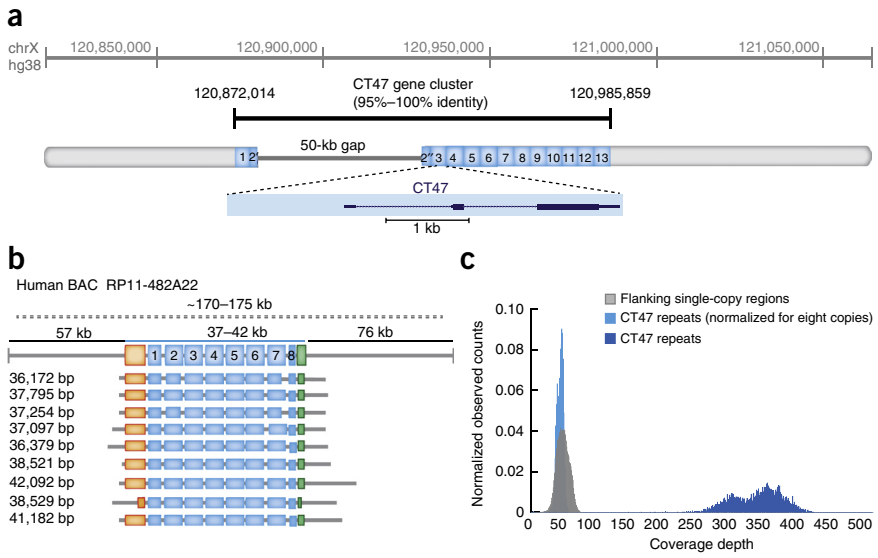


Figure 6 | Resolution of CT47 repeat copy-number estimate on human chromosome Xq24. (**a**) BAC end sequence alignments (RP11-482A22: AQ630638 and AZ517599) span a 247-kb region, including 13 annotated CT47 genes[16] (each within a 4.8-kb tandem repeat), and a 50-kb scaffold gap in the GRCh38/hg38 reference assembly. (**b**) Nine MinION reads from high–molecular weight BAC DNA span the length of the CT47-repeat region, providing evidence for eight tandem copies of the repeat. Insert size estimated from pulse-field gel electrophoresis (dashed line) with flanking regions (black lines) and repeat region (blue line) are shown. Single-copy regions before and after the repeats are shown in orange (6.6 kb) and green (2.6 kb), respectively, along with repeat copies (blue) and read alignment in flanking regions (gray). The size of each read is shown to its left. (**c**) Shearing BAC DNA to increase sequence coverage provided copy-number estimates by read depth. All bases not included in the CT47 repeat unit are labeled as flanking regions (gray distribution; mean: 46.2-base coverage). Base coverage across the CT47 repeats was summarized over one copy of the repeat to provide an estimate of the combined number (dark blue distribution; mean: 329.3-base coverage) and was similar to single-copy estimates when normalized for eight copies (light blue distribution; mean: 41.15-base coverage).

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0004

longer reads, but at a lower frequency, when very large intact DNA fragments were delivered to the sequencer (for example, a full-length 48-kb 2D read of phage λ DNA mapped back to the reference with 87% identity (**Supplementary Fig. 13**)). We reasoned that long MinION reads, coupled with our high-confidence alignment strategy, could be used to resolve complex and often unfinished regions of genomes.

To test this, we examined the organization of a human-specific tandem-repeat cluster spanning a putative 50-kb assembly gap on human Xq24 (hg38 chrX:120,814,747–121,061,920) (**Fig. 6a**). Each 4,861-bp tandem repeat in this region contains a single annotated cancer-testis gene from the CT47 gene family with observed expression in testes, lung and esophageal cancer cells[16]. The high level of homology between adjacent copies (95%–100% sequence identity) is likely to result in recombination or replication errors, leading to alleles with different numbers of repeats that are often difficult to represent accurately by standard short-read assembly[17]. Furthermore, copy-number expansion and contraction involving genes contribute to variability in gene expression, epigenetic regulation and association with human disease[18,19].

We used the MinION to acquire very long reads from a human BAC (RP11-482A22) that contained the CT47 repeats within the unresolved Xq24 segment. Nine 2D reads from 36 kb to 42 kb spanned all the repeats and together indicated eight tandem copies within the gap (Online Methods, **Fig. 6b** and **Supplementary Data**). This copy-number prediction was supported by pulse-field gel electrophoresis, which revealed a repeat array of 37–42 kb, or 7.5–8.6 copies of the 4.8-kb repeat (**Supplementary Fig. 14**). As an additional test, we obtained 40×–60× sequence coverage of the unresolved Xq24 segment using shorter (~10 kb) MinION reads from sheared BAC DNA. A copy-number estimate based on these reads also indicated eight CT47 repeats within the unresolved region (**Fig. 6c**).

DISCUSSION

We began this study by documenting MinION performance using M13 phage dsDNA. We found that consecutive reads of adaptor-linked template and complement DNA strands (~14.4 kb total) were routinely achieved. Approximately 99% of 2D reads mapped to a reference (M13 or phage λ DNA control) and yielded 85% average identity. Using expectation-maximization training of an HMM, we were able to robustly parse the error sources into mismatches, insertions and deletions. This information was used to generate high-confidence alignments that allowed us to call SNVs accurately and characterize an unresolved region of human Xq24 rich in repetitive DNA. A dual–MinION sequencing strategy that employed both long-read scaffolds and higher-coverage shorter reads was essential for copy-number estimates in that region.

Comparisons with prior results[11,20] demonstrated improved read quality during MAP. We anticipate that the number of correct base calls will continue to increase beyond the average 85% identity observed in the current study. We also expect that the MinION will be used to report features of genomic DNA that are observable because the nanopore sensor directly touches each base on native DNA strands. These features include epigenetic modifications[21–23], abasic residues[24,25], DNA adducts[26], thymine-thymine dimers and strand breaks.

In summary, we have shown that the MinION has sufficient accuracy to resolve important biological questions by sequencing long, native DNA strands. This accuracy is improving rapidly.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. ENA: PRJEB8230, ERP009289.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Human Genome Research Institute of the US National Institutes of Health under award numbers HG006321 (M.A.), HG007827 (M.A.) and U54HG007990 (B.P.). The authors thank Oxford Nanopore Technologies for their gift to the UCSC Nanopore Group. The authors also thank D. Deamer for support, reading of the manuscript and helpful discussion. The authors gratefully acknowledge D. Haussler and J. Kent for their support.

AUTHOR CONTRIBUTIONS

M.A. conceived experiments and directed research. B.P. conceived and directed bioinformatics analysis. B.P., M.J., I.T.F. and K.H.M. were responsible for bioinformatics analysis and software development. M.J. and H.E.O. were responsible for the completion of sequencing experiments and data processing. M.J. and H.E.O. were responsible for preparing DNA sequencing standards. H.E.O. was responsible for Sanger sequencing of M13 dsDNA. B.P. and I.T.F. were responsible for *k*-mer and BLAST analysis. B.P. and M.J. were responsible for SNV analysis. B.P. developed and implemented expectation-maximization and realignment strategies. K.H.M. conceived and directed BAC experiments and data analysis. All authors contributed to the writing, editing and completion of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests. Details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
2. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/pdf/1303.3997.pdf (2013).
3. Frith, M.C., Wan, R. & Horton, P. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res.* **38**, e100 (2010).
4. Harris, R.S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).
5. Benson, D.A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
7. Do, C.B. & Batzoglou, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **26**, 897–899 (2008).
8. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
9. Schwartz, A.S. & Pachter, L. Multiple alignment by sequence annealing. *Bioinformatics* **23**, e24–e29 (2007).
10. Quick, J., Quinlan, A. & Loman, N. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).
11. Ashton, P.M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* doi:10.1038/nbt.3103 (8 December 2014).
12. Davey, J.W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0005

13. Bourlat, S.J. *et al.* Genomics in marine monitoring: new opportunities for assessing marine health status. *Mar. Pollut. Bull.* **74**, 19–31 (2013).
14. Stucki, D. & Gagneux, S. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. *Tuberculosis (Edinb.)* **93**, 30–39 (2013).
15. Holmes, I. & Bruno, W.J. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17**, 803–820 (2001).
16. Chen, Y.T., Iseli, C. & Venditti, C. Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes Chromosomes Cancer* **45**, 392–400 (2006).
17. Treangen, T.J. & Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
18. Tremblay, D.C., Alexander, G., Moseley, S. & Chadwick, B.P. Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics* **11**, 632 (2010).
19. Brahmachary, M. *et al.* Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS Genet.* **10**, e1004418 (2014).
20. Mikheyev, A.S. & Tin, M.M. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).
21. Schreiber, J. *et al.* Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc. Natl. Acad. Sci. USA* **110**, 18910–18915 (2013).
22. Laszlo, A.H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. USA* **110**, 18904–18909 (2013).
23. Wescoe, Z.L., Schreiber, J. & Akeson, M. Nanopores discriminate among five C5-cytosine variants in DNA. *J. Am. Chem. Soc.* **136**, 16582–16587 (2014).
24. Cherf, G.M. *et al.* Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.* **30**, 344–348 (2012).
25. Lieberman, K.R., Dahl, J.M., Mai, A.H., Akeson, M. & Wang, H. Dynamics of the translocation step measured in individual DNA polymerase complexes. *J. Am. Chem. Soc.* **134**, 18816–18823 (2012).
26. Schibel, A.E. *et al.* Nanopore detection of 8-oxo-7,8-dihydro-2′-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site. *J. Am. Chem. Soc.* **132**, 17992–17995 (2010).

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0006

ONLINE METHODS

M13 MinION experiments. We generated three replicate experiments with M13mp18 phage dsDNA to establish the reproducibility and performance characteristics of the MinION. Below we describe the M13 sequencing-standard preparation and MinION sequencing protocols.

M13mp18 DNA sequencing standard. M13mp18 dsDNA was obtained from New England Biolabs (NEB) (catalog no. N4018S). The host for this phage is *E. coli* strain ER2738, and the genome is 7.2 kb in size with a 42% average GC content. Thirty micrograms of M13mp18 was linearized by means of overnight double digestion with High-Fidelity HindIII (NEB, catalog no. R3104S) and High-Fidelity BamHI (NEB, catalog no. R3136S). Digests were performed according to NEB recommendations using Cut Smart Buffer supplied with restriction enzymes. Two hundred nanograms of M13mp18 double digest was run on a 1% Tris borate EDTA (TBE) agarose gel to confirm complete linearization of the circular replicative-form genome. The restriction digest was then extracted once with an equal volume of TE buffer (10 mM Tris, 1 mM EDTA, pH 8)-equilibrated phenol:chloroform (OmniPur, catalog no. 6805) and twice with TE buffer–equilibrated chloroform, pH 8, and then ethanol precipitated by the addition of 1/10 volume of 3 M sodium acetate, pH 5.2 (Teknova, catalog no. S0296), and 2 volumes of ice-cold 100% ethanol. Samples were centrifuged to pellet DNA, and the M13mp18 pellet was washed twice with 70% ethanol, allowed to dry to remove ethanol, resuspended in MilliQ water and quantitated using a Nanodrop. The M13 sequence was confirmed using Sanger sequencing (UC Berkeley DNA Sequencing Facility, with an ABI Model 3730 XL DNA Sequencer (Applied Biosystems, Life Technologies, Thermo Fisher Scientific)). Sequencing primers TAAGGTAATTCACAATGATTAAAGTTG, CTGTGGAATGCTACAGGC, CACCTTTAATGAATAATTTCCGTC, CATGCTCGTAAATTAGGATGG, GTTTTACGTGCTAATAATTTTGATATG, CAAGGCCGATAGTTTGAGT, CACTGGCCGTCGTTTTA, GAGGCTTTATTGCTTAATTTTGC, AGGTCTTTACCCTGACTATTATAG, AGGCTTTGAGGACTAAAGAC, AATGGATCTTCATTAAAGCCAG, CAGCCTTTACAGAGAGAATAAC, TCCGGCTTAGGTTGGG, GTGAGGCGGTCAGTATTAAC, GAGATAGGGTTGAGTGTTGT and TTCTCCGTGGGAACAAAC were obtained from Integrated DNA Technologies (http://www.idt.com/).

M13 MinION sequencing. The libraries for MinION runs were prepared as recommended by ONT. Unsheared DNA was used for preparation of the M13 sequencing library. For BAC DNA, sequencing libraries were prepared using unsheared DNA as well as DNA sheared to an average length of 10 kb using g-TUBE (Covaris, catalog no. 520079). Briefly, the DNA sample was spiked with ONT λ DNA control, end-repaired using NEBNext End Repair Module (NEB, catalog no. E6050S) and cleaned up using Agencourt AMPure XP beads (Beckman Coulter, catalog no. A63880). The purified end-repaired DNA then underwent dA tailing with the NEB dA-Tailing Module (NEB, catalog no. E6053S). This was followed by ligation of ONT sequencing adaptors (adaptor Mix and HP adaptor) using Blunt/TA Ligase Master Mix (NEB, catalog no. M0367S). Using Dynabeads His-Tag Isolation and Pulldown (Life Technologies, catalog no. 10103D), we enriched the library for DNA molecules ligated to the ONT HP adaptor. The adapted and enriched DNA was eluted in ONT-supplied elution buffer. This prepared library was then mixed with proprietary ONT EP Buffer and ONT Fuel Mix before being added to the MinION flow cell. Three 48-h sequencing runs were performed, each using a new flow cell.

The MinION data were base called using ONT Metrichor software (workflow R7.X 2D rev1.9). The base caller used classifies reads as pass or fail. Unless otherwise noted, all the analyses reported in this paper were performed using the 'pass' reads from R7.3 chemistry.

Establishing a mapping strategy for MinION reads. We experimented with four different initial read-mapping programs: BLASR[1] (PacBio's long-read mapper designed for mapping PacBio reads; commit abf9c38c55c2fb5f 40316885dce39f5308c9ff25 from https://github.com/PacificBiosciences/blasr), BWA-MEM Release 0.7.11 (refs. 2,27) (H. Li's popular adaptation of the BWA mapper altered for handling long reads), LAST Version 490 (refs. 3,28) (a fast, sensitive, adaptable and popular pairwise-alignment tool) and LASTZ Release 1.02.00 (ref. 4) (a more traditional BLAST-type seed-and-extend program).

For each mapping experiment, reads were mapped both to the M13 reference sequence and to control DNA, a 3.8-kb segment of λ phage DNA supplied by ONT to be used in each experiment to measure baseline performance. For each mapping program, a sizable fraction of reads could not be aligned to either reference when the default parameters were used (data not shown). The use of tuned parameters substantially improved the number of reads mapped to the reference sequences.

To establish whether the mappers produced substantial numbers of false positive mappings, the reference sequences were reversed but not complemented, and the reads were mapped to these reversed sequences. The rationale for this experiment was that in the resulting reversed sequences, the base composition in terms of GC content and reversible Markov chain–like properties would be preserved, but it was highly unlikely that the sequences would be similar to the reads (**Supplementary Fig. 3**).

BLAST analysis for unmapped reads. In order to characterize the small minority of unmapped reads, we used BLAST 2.2.29 to align the unmapped reads to the NCBI Nucleotide database. The Nucleotide database contains entries from all of the traditional divisions of GenBank, the European Molecular Biology Laboratory and the DNA Data Bank of Japan[5,6]. The majority of unmapped 2D reads had BLAST hits (**Fig. 2** and **Supplementary Table 3**), most representing a low level of *E. coli* contamination.

Learning the MinION error model. The MinION error model we propose is a five-state pair HMM[29] that has two sets of insertion-deletion states (**Supplementary Fig. 4**), one set for modeling short insertions and deletions and one for modeling long insertions and deletions. The latter was included to account for large gaps at the beginnings and ends of the alignments—that is, to convert a local alignment model into a global alignment, as described by Durbin *et al.*[29]. To train the model we used a hybrid form of the Baum-Welch algorithm (a form of expectation maximization) that, for speed, works within an alignment band around a fixed guide alignment[8] for each read, with the guide alignments provided by a mapping program and the band constructed as

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0007

described by Paten *et al.*[8], using C code adapted from the Cactus alignment program[30]. In contrast to alignment models learned from sequences related by evolution, no assumption of reversibility (and therefore symmetry) was made, and parameters for each transition and emission were learned independently.

We trained the alignment model for each possible combination of guide mapping program (tuned versions of the four mapping programs tested), MinION run (of three replicates) and read-type set (template, complement and 2D). For each training experiment we performed three independent runs, in each case starting from a randomly parameterized model and running for 100 iterations. **Supplementary Figure 5** shows the results of one training experiment, in which there is convergence of log-likelihood for all three runs to essentially the same value. **Supplementary Figure 5** also shows the resulting transition parameters for each read type; we observed excellent agreement in parameter estimates both between runs for the same training experiment and between training experiments with different MinION runs and different guide alignments, indicating that our parameter estimates were robust.

Figure 3a,b shows, as a cross-check, the calculation of insertion, deletion and substitution rates for 2D reads from realignments computed (see below) from each guide alignment using the alignment and the trained model. In each case, despite the fact that the starting guide alignments had very different estimates of these error rates, the realigned alignments gave consistently close error rates for these parameters. Interestingly, these values agreed relatively closely with the starting tuned-BLASR alignments, indicating it was most closely parameterized to our estimates of the maximum-likelihood rates.

Realignment with a trained model. For each possible combination of guide mapping program (tuned versions of BLASR, BWA-MEM, LAST and LASTZ; see **Supplementary Table 2**), MinION run (of three replicates) and read-type set (template, complement and 2D), we trained the alignment model and then realigned the reads using the resulting model. We call such alignments trained realignments. To realign the reads we used the aforementioned banding strategy around the guide alignment and picked a single alignment using the AMAP objective function[9], which calculates an alignment that accounts for the posterior expectation of each match and indel. As a control experiment to account for the effects of realigning the reads, we also realigned the reads using the same guide-alignment strategy and objective function, but with an untrained model, the default HMM used by Cactus[30], which was parameterized for vertebrate sequences related by natural selection. The control experiment showed that such alignments had substantially lower identity, indicating that the training, and not the process of realignment, was responsible for the improvement in identity (**Supplementary Fig. 9**).

SNV calling with the MinION. To determine how useful MinION reads are for simple SNV discovery in monoploid genomes, we took the M13mp18 reference sequence and randomly introduced substitutions at frequencies of 1%, 5%, 10% and 20%, picking the alternate allele with equal probability for each possible alternate base. We called each altered sequence a mutated reference sequence. For each read type for each replicate of the M13mp18 experiment, we aligned the reads to each mutated reference sequence with a given mapper and ran an algorithm to call SNVs with respect to the mutated reference sequence.

Briefly, the SNV-calling algorithm (see **Supplementary Note 4** for a full description) has two steps: computing posterior alignment match probabilities between the bases in the reads and the reference, and calculating posterior base-calling probabilities for each reference base. By varying the threshold on the posterior base-calling probability, we traded precision for recall (**Fig. 5**). The reported precision and recall values were chosen to optimize the overall *F*-score.

The posterior match probabilities were computed using the guided-realignment strategy described above. The HMM used was composed by combining the described pair HMM (trained using expectation maximization on 2D reads with tuned LAST used as the guide alignment, as described earlier) with a substitution model that accounts for the introduced mismatches. Each model was described as a branch transducer[15], and the models were combined to create an overall HMM, using the evolutionary HMM formalism[15]. The addition of the substitution model was found to be essential for high performance; **Supplementary Note 4** describes the parameters used and algorithm variations.

Sequence scaffolding across the CT47 repeat cluster. High–molecular weight BAC DNA (RP11-482A22) was isolated using standard methods for purification of large constructs (QIAGEN Large-Construct Kit, catalog no. 12462). To avoid DNA shearing for high–molecular weight sequencing, we performed NotI-HF (NEB, catalog no. R3189S) restriction digestion (expected to isolate the insert from pBACe3.6 cloning vector, gi|4878025) followed by end repair using Klenow in the same mix. This mixture underwent dA tailing directly after being added with separately end-repaired ONT-supplied control DNA, and the rest of the steps then proceeded according to the standard ONT recommendations, as mentioned above. The device was operated using ONT's MinKNOW software according to the provided instructions. The flow cells used were chemistry version R6.0 and R7.0. The read files were base called using ONT's Metrichor software, version 2D base calling, v1.2 and v1.3.1.

Long reads spanning the CT47-repeat cluster were identified using three sequence models[31]: a single-copy sequence directly upstream of the repeat array (6.6 kb, hg38 chrX:120865735 120872351), the CT47 repeat (4.8 kb, hg38 chrX:120932375-120937233) and a single-copy sequence directly downstream from the repeat array (2.7 kb, hg38 chrX:120986928-120989651). Reads were trimmed to the only present sequences involved in the repeat-classification models, and Pecan software[8] was used to generate multiple alignment of reads (data available in the European Nucleotide Archive; the primary accession number is PRJEB8230, and the secondary accession number is ERP009289).

Copy-number estimates by sheared BAC sequencing. To increase the MinION sequence throughput, we sheared RP11-482A22 BAC DNA to an average fragment length of 10 kb using g-TUBE (Covaris, catalog no. 520079). By alignment to the hg38 reference sequence (hg38 chrX:120,814,747–121,061,920, omitting a 50-kb scaffold gap), using tuned BLASR (as described above), we identified 2,006 2D reads that mapped to the RP11-482A22 DNA. Base coverage was determined from a sorted-alignment RP11-482A22 BAM file using bedtools genomecov[32] with the

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0008

doi:10.1038/nmeth.3290

command bedtools genomecov -d -ibam mapping.sorted.bam. Coverage estimates were converted to a BED file with each row entry defining coverage at a single base and at base + 1, and then they were subdivided into bases that overlapped with the CT47 repeat region and those that did not, with the latter labeled as flanking regions (bedtools intersect -woa and -v, respectively)[32]. A histogram of base coverage was generated to encompass all flanking bases and was determined to have a mean coverage value of 46.2 bases. Base-coverage estimates across the CT47 repeats were merged to represent a combined coverage depth over a single 4.8-kb repeat unit (mean observed base coverage: 329.3). Normalization of the read depth for eight copies of the repeat predicted an average read depth of 41 bases. We obtained the distribution of the normalized read depth by dividing by 8 across all base positions of the repeat with combined sequence depth.

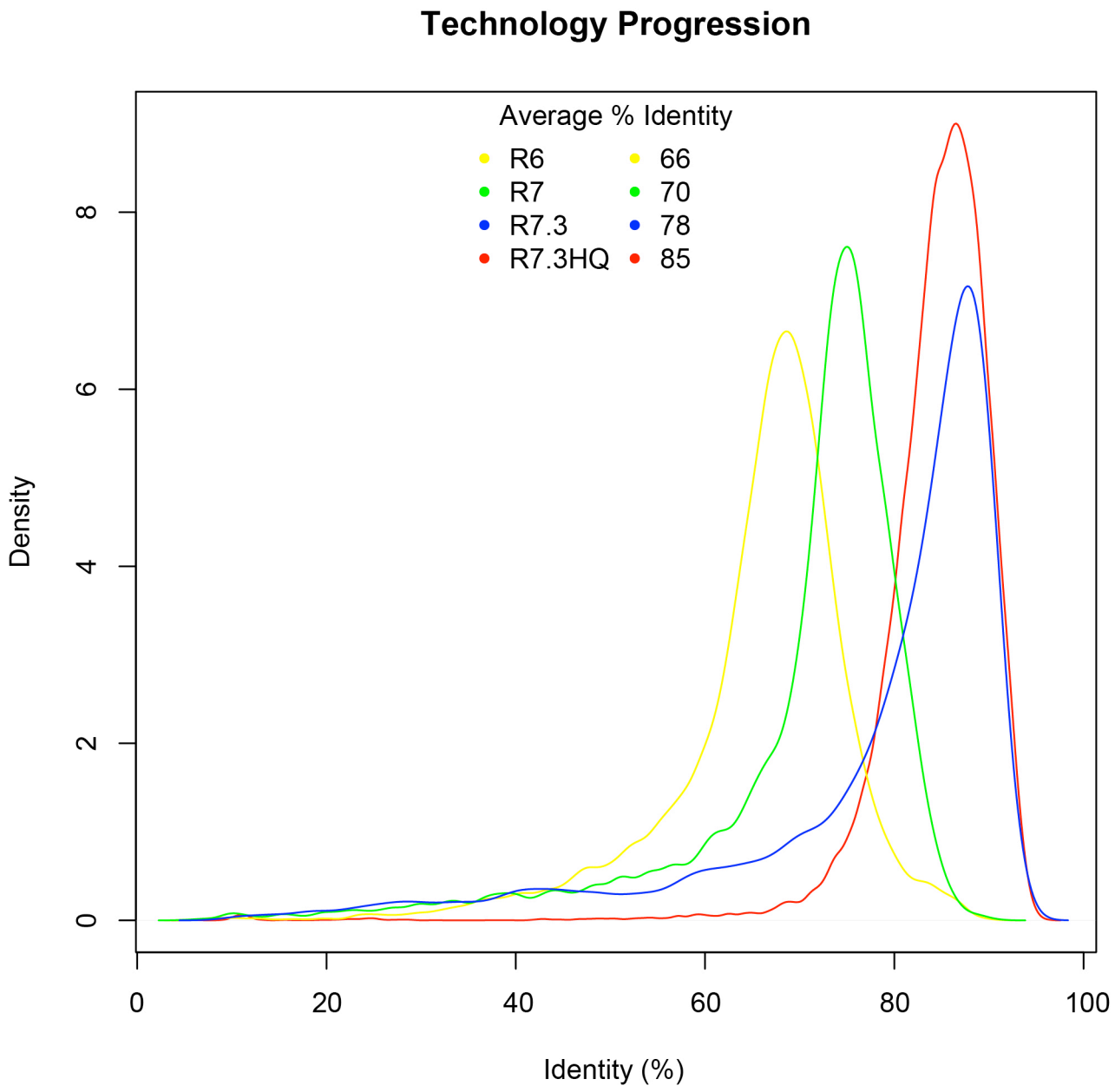
Pulse-field gel electrophoresis validation. The RP11-482A22 BAC insert length estimate of NotI-HF-digested (NEB, catalog no. R3189S) or AatII-digested (NEB, catalog no. R0117S) DNA (1 µg) was determined by pulse-field gel electrophoresis (PFGE) using a CHEF-DRII system (Bio-Rad). Length estimates were determined using standard PFGE markers Low-range (NEB, catalog no. N0350S) and MidRange I (NEB, catalog no. NE551S). Samples were run for 15 h (gradient, 6.0 V/cm; angle, 120°; switch time, linear; initial ramping, 0.2 s, finishing at 26 s) in 1% Pulsed Field Certified Agarose (Bio-Rad) and 0.5× TBE buffer at 4 °C. Banding was identified using standard SYBR Gold (Life Technologies) staining.

Code availability. The analysis software is open source and is available online (nanopore pipeline at https://github.com/mitenjain/nanopore and marginAlign pipeline at https://github.com/benedictpaten/marginAlign).

27. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM https://github.com/lh3/bwa/blob/master/NEWS.md/#release-079-19-may-2014 (2014).
28. Frith, M.C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80 (2010).
29. Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. (The Press Syndicate of The University of Cambridge, 1998).
30. Paten, B. *et al.* Cactus: algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
31. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
32. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

© 2015 Nature America, Inc. All rights reserved.

QIAGEN Exhibit 1028.0009



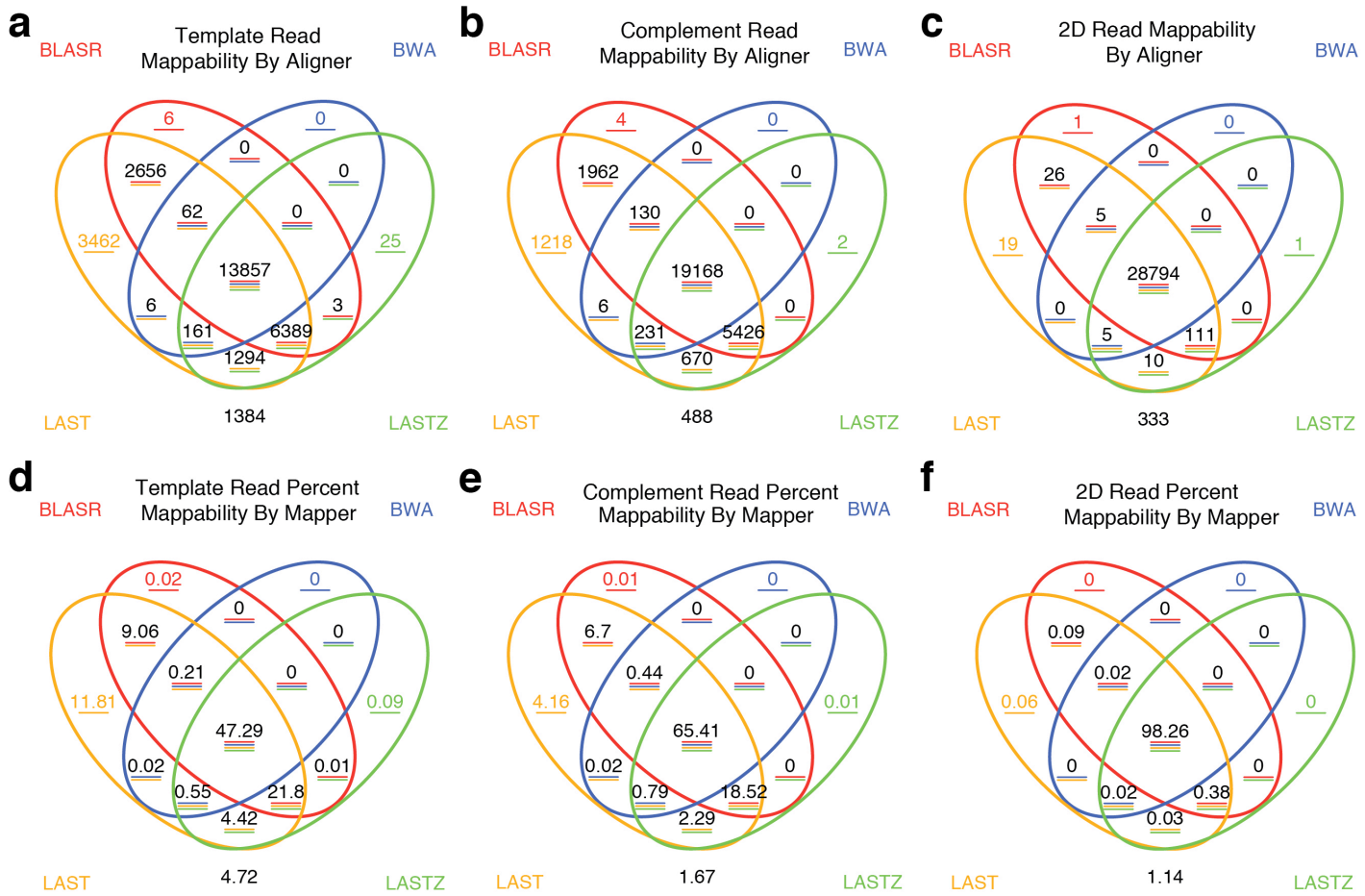
Supplementary Figure 1

MinION technology progression.

Progression of read identity distributions with MinION versions since June 2014.

QIAGEN Exhibit 1028.0010

Nature Methods: doi:10.1038/nmeth.3290

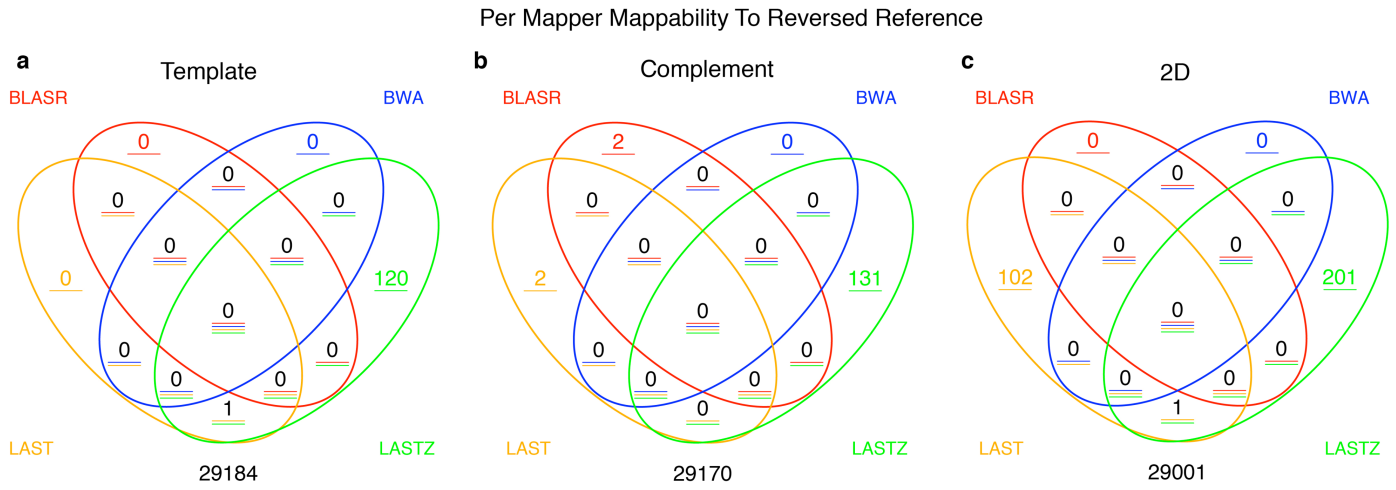


Supplementary Figure 2

Venn diagram representing read mappability for MinION reads across three replicate M13 experiments using R7.3 chemistry.

Mappability represents the proportion of reads that can be aligned to either the M13 or the phage λ DNA control using the tuned parameters for each mapper. In our analysis, 2D reads had the highest mappability, with 99% of reads being mappable, followed by complement and template reads, with 98% and 95% of their respective read proportions being mappable. Among the four aligners used, LAST and LASTZ performed the best for M13, with LAST capturing the greatest proportion of mappable reads on its own.

QIAGEN Exhibit 1028.0011

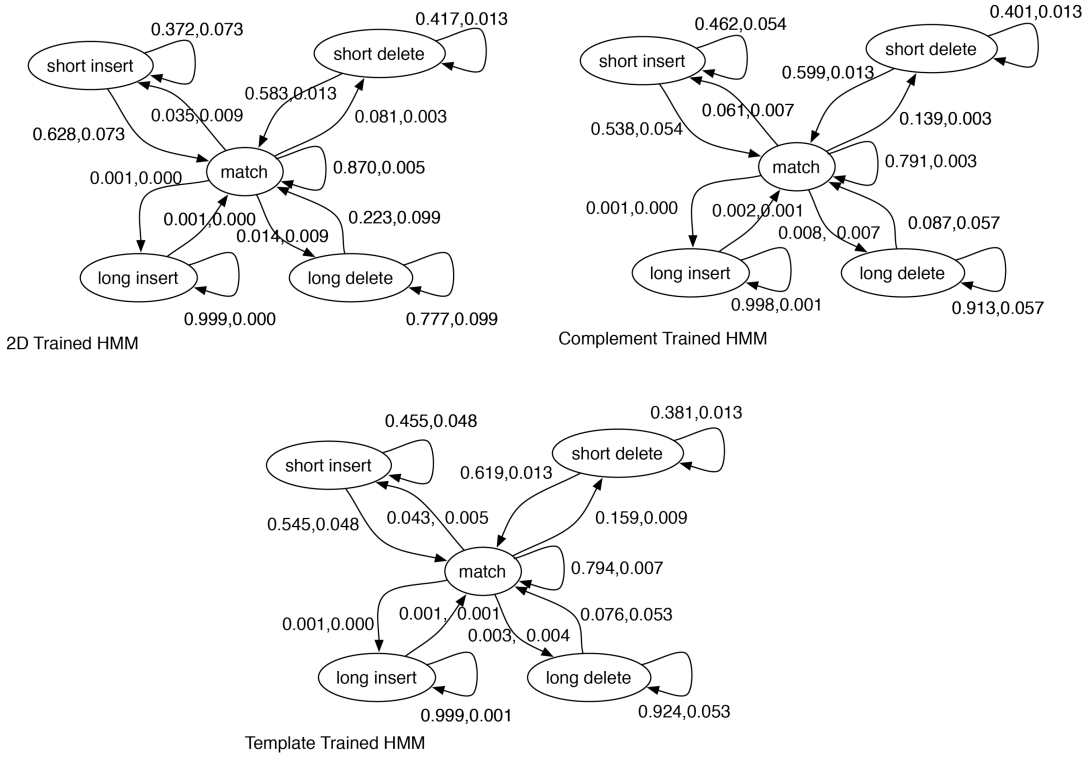


Supplementary Figure 3

Venn diagram representing read mappability to a reversed reference for MinION reads from three replicate M13 experiments using R7.3 chemistry.

Because the reference was reversed, effectively no reads should map; this is thus a proxy measure of specificity. Results were obtained using the tuned alignment parameters.

QIAGEN Exhibit 1028.0012



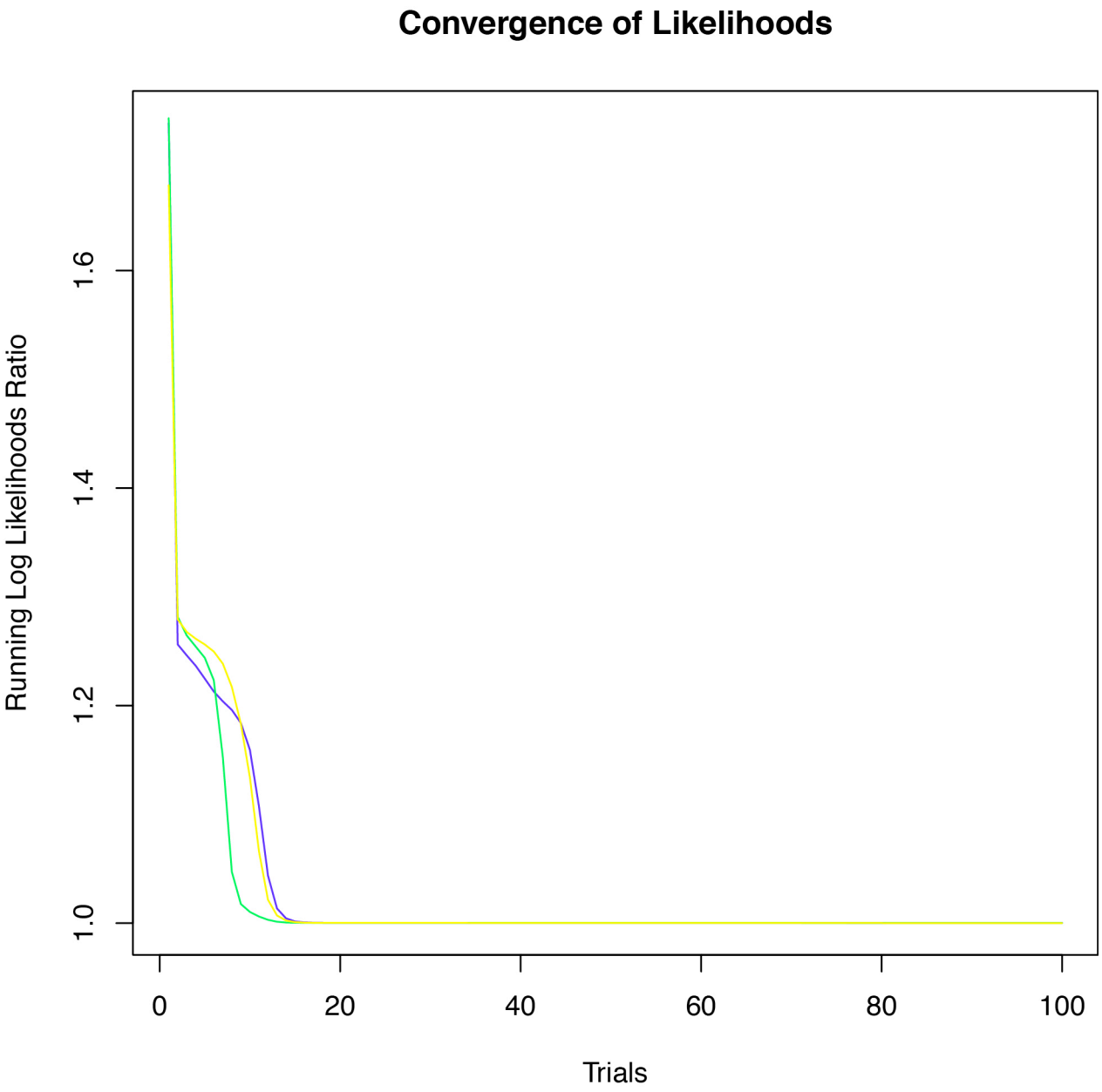
Supplementary Figure 4

Structure for the hidden Markov model (HMM) used for expectation maximization (EM).

Structure of HMM used for EM, along with the estimated parameters for transition probabilities for template, complement and 2D reads. For each transition in order, the mean estimates and standard error across all experiments for that read type are shown.

QIAGEN Exhibit 1028.0013

Nature Methods: doi:10.1038/nmeth.3290



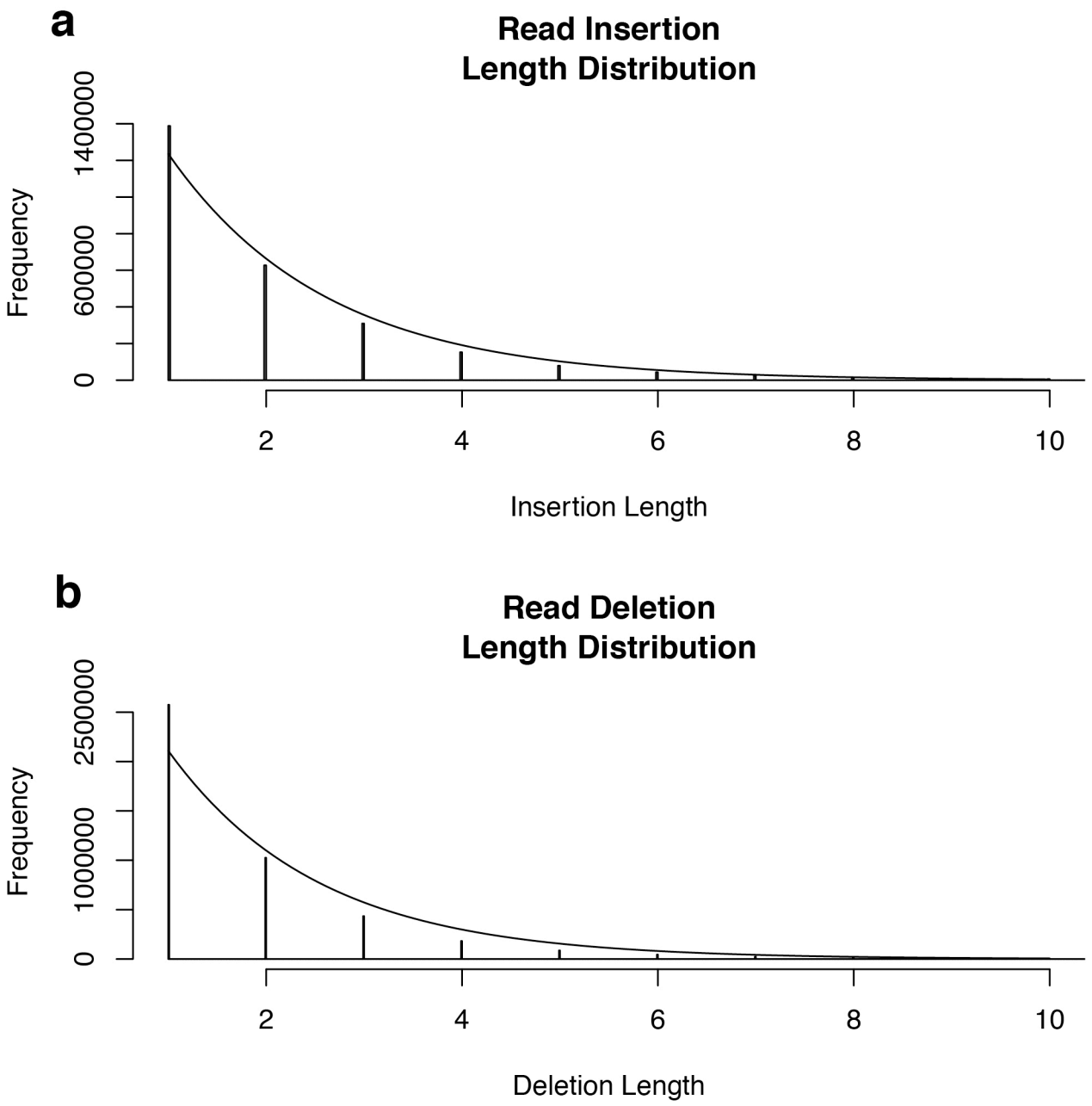
Supplementary Figure 5

Convergence of log-likelihood ratios achieved using expectation maximization.

Convergences of log-likelihood for three independent runs of expectation maximization, each from a randomly parameterized model, each run for 100 iterations of training. The *y*-axis gives likelihood normalized by the highest log-likelihood found. The training used 2D reads from one MinION run of the M13 data using release R7.3 chemistry and a guide alignment generated by tuned LAST.

QIAGEN Exhibit 1028.0014

Nature Methods: doi:10.1038/nmeth.3290

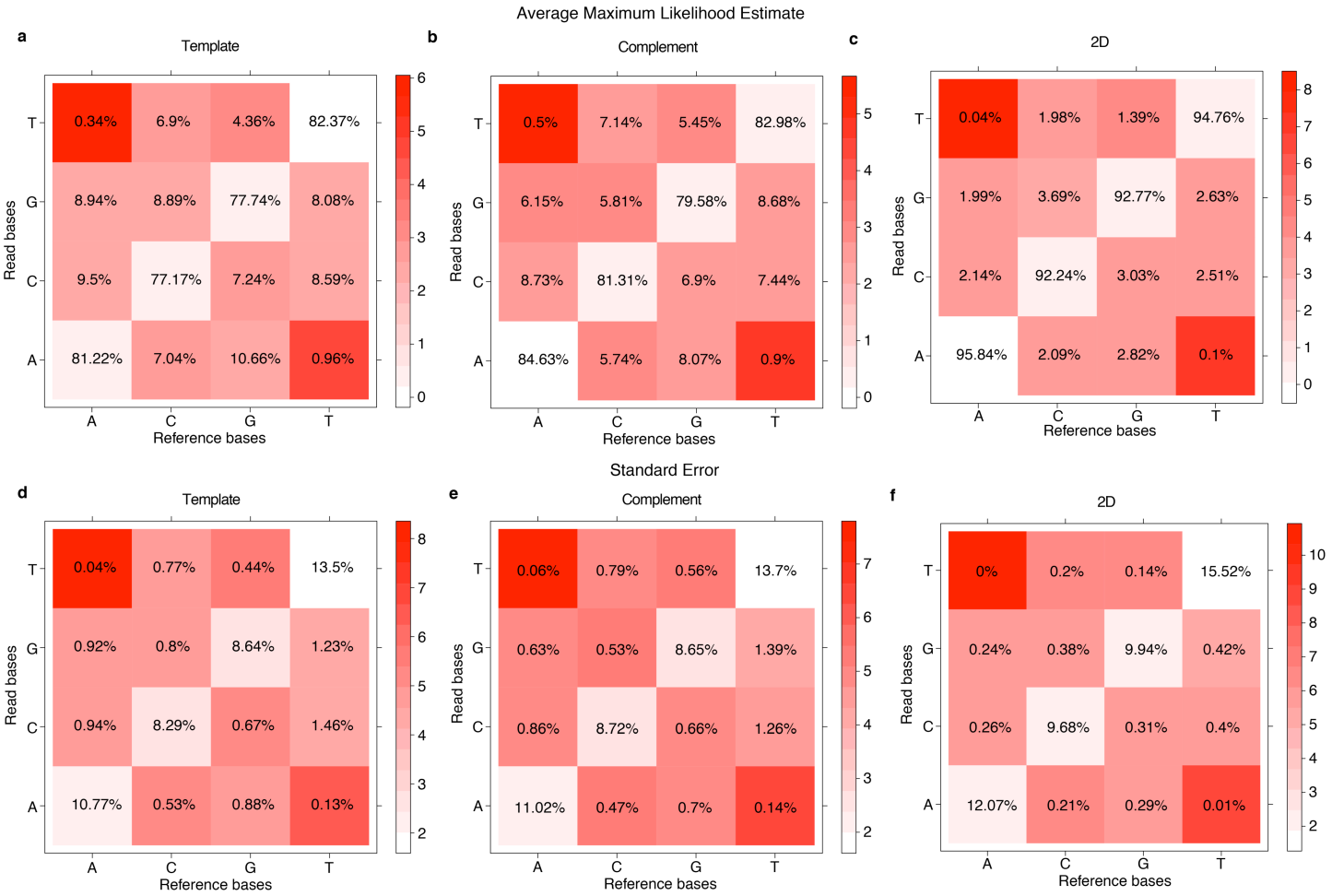


Supplementary Figure 6

Frequency plots for insertions and deletions in MinION read alignments.

Representative insertion and deletion plot for reads (fitted with an exponential distribution) from one M13 experiment using R7.3 chemistry, aligned using expectation maximization–trained LAST.

QIAGEN Exhibit 1028.0015

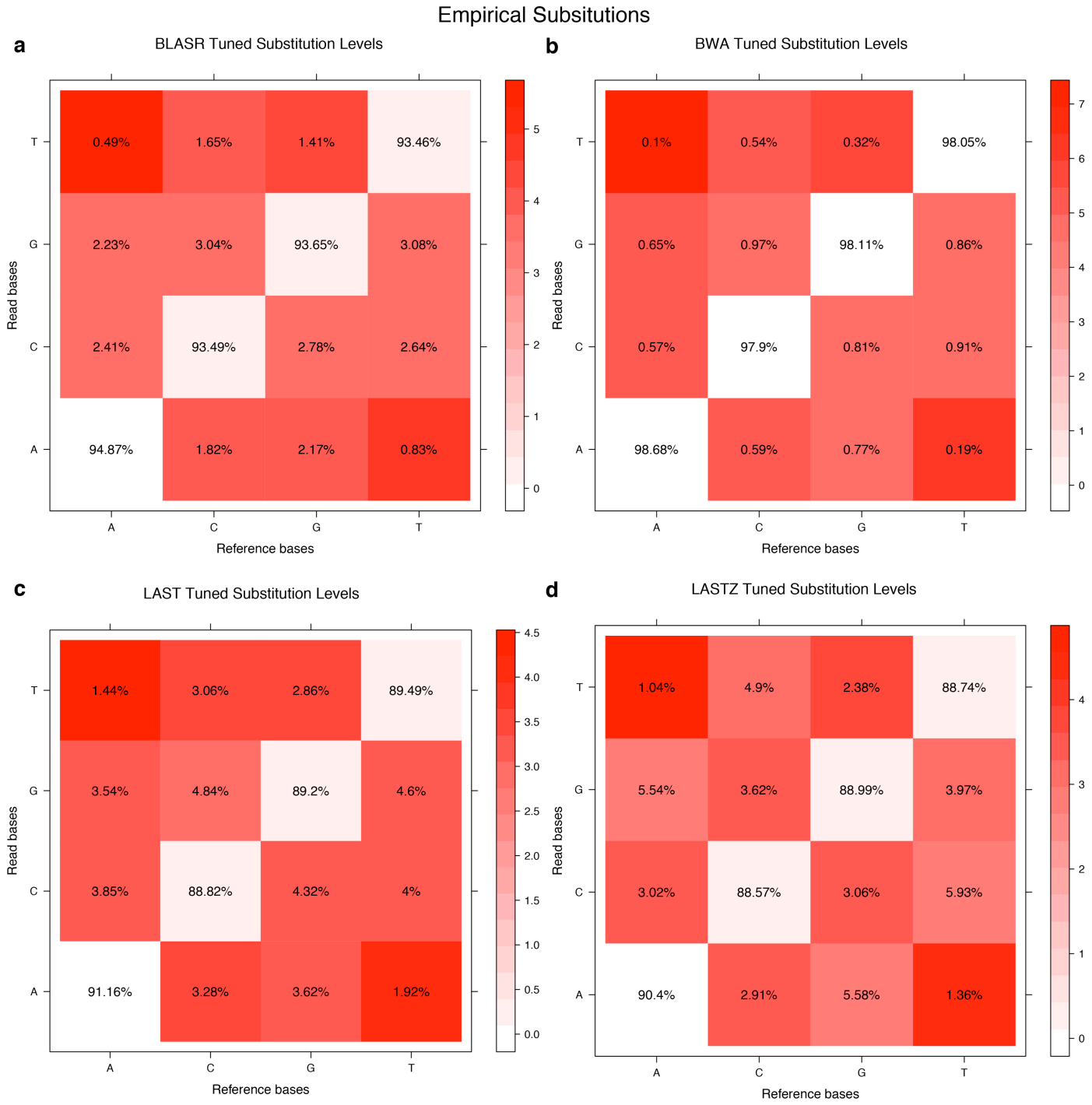


Supplementary Figure 7

Substitution matrices from alignments using expectation maximization–trained model.

Maximum-likelihood estimates and standard-error parameters for substitution matrices show trends across template, complement and 2D reads across three M13 experiments using R7.3 chemistry. The top row illustrates the average maximum-likelihood estimates for these substitutions, with the standard error represented in the lower row. For all aligners, thymine-to-adenosine and adenosine-to-thymine substitution rates were low, indicating that the device rarely miscalled one as the other. The color scheme is fitted on a log scale, and the substitution values are on an absolute scale.

QIAGEN Exhibit 1028.0016

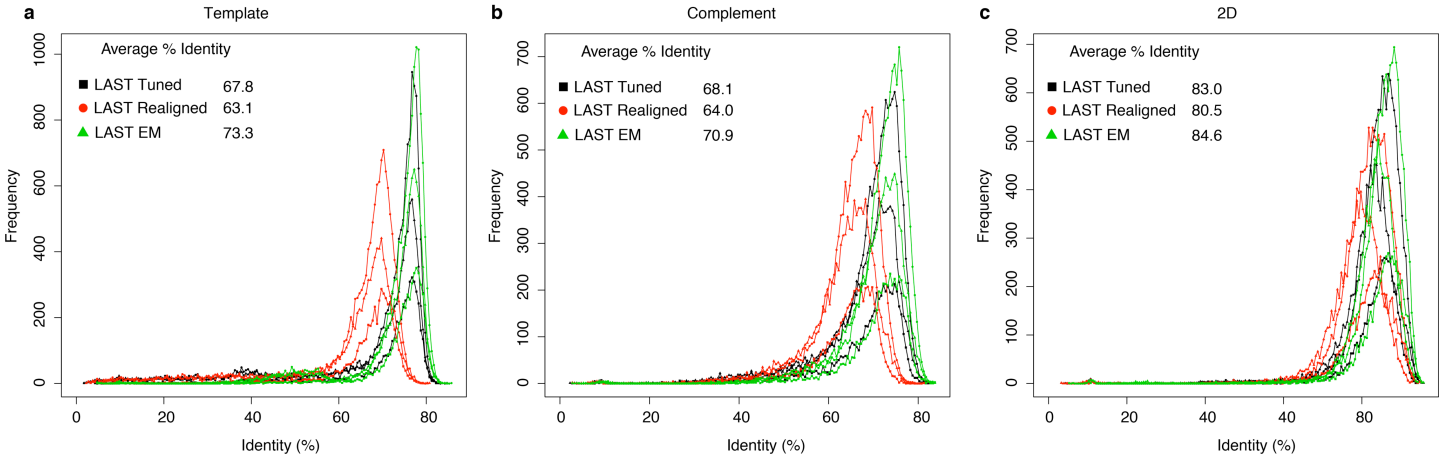


Supplementary Figure 8

Substitution matrices from alignments using tuned parameters.

Substitution matrices for each of the four tuned aligners across three M13 experiments using R7.3 chemistry. For all aligners, thymine-to-adenosine and adenosine-to-thymine substitution rates were low, indicating that the device rarely miscalled one as the other. The color scheme is fitted on a log scale, and the substitution values are on an absolute scale.

QIAGEN Exhibit 1028.0017

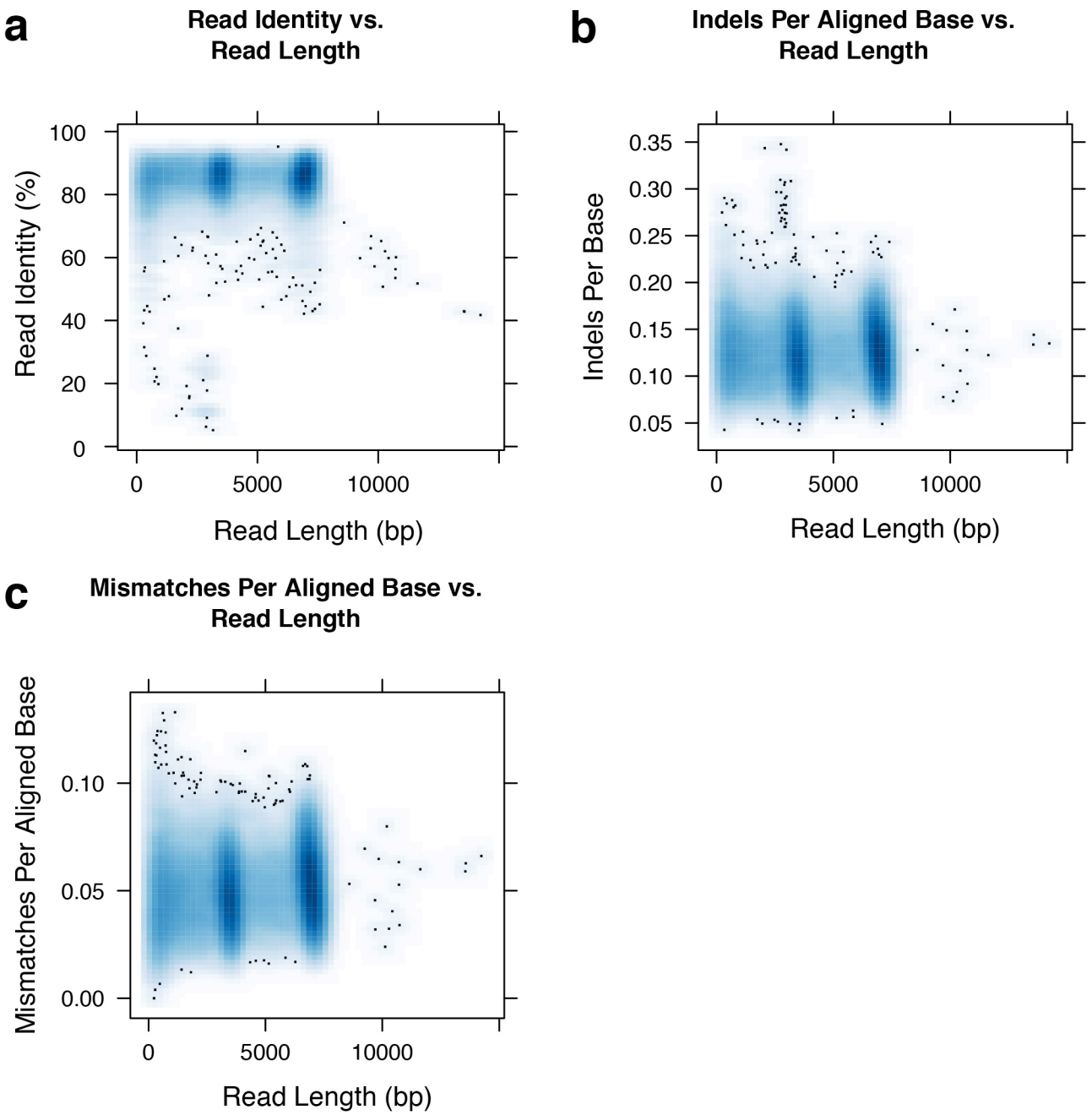


Supplementary Figure 9

Realignment improves read identity.

Read identity for template, complement and 2D reads for three M13 replicate experiments using R7.3 chemistry, aligned using LAST. Three versions of the LAST alignment are shown: tuned LAST, trained LAST realignments and naive LAST realignments.

QIAGEN Exhibit 1028.0018

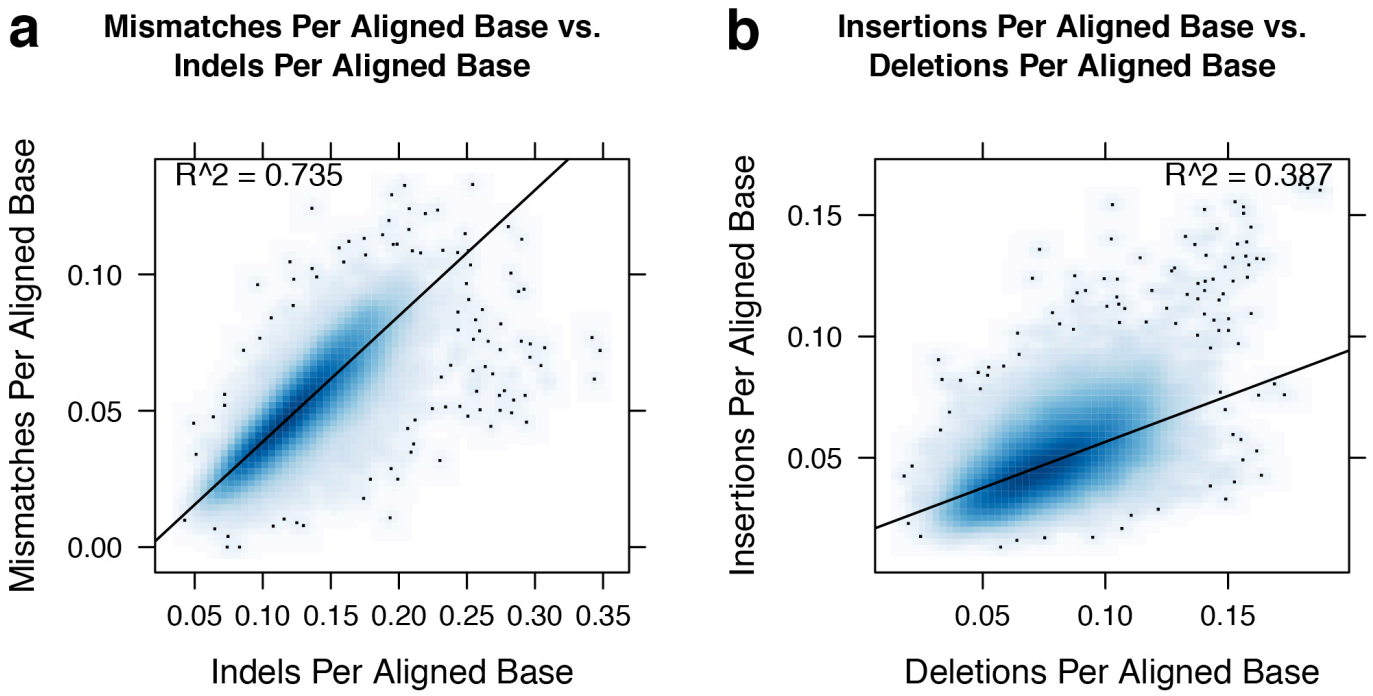


Supplementary Figure 10

An alignment quality measurement for 2D reads across three M13 replicate experiments.

Alignments were obtained using expectation maximization–trained LAST realignments. The two density clusters correspond to M13 and phage λ DNA control.

QIAGEN Exhibit 1028.0019

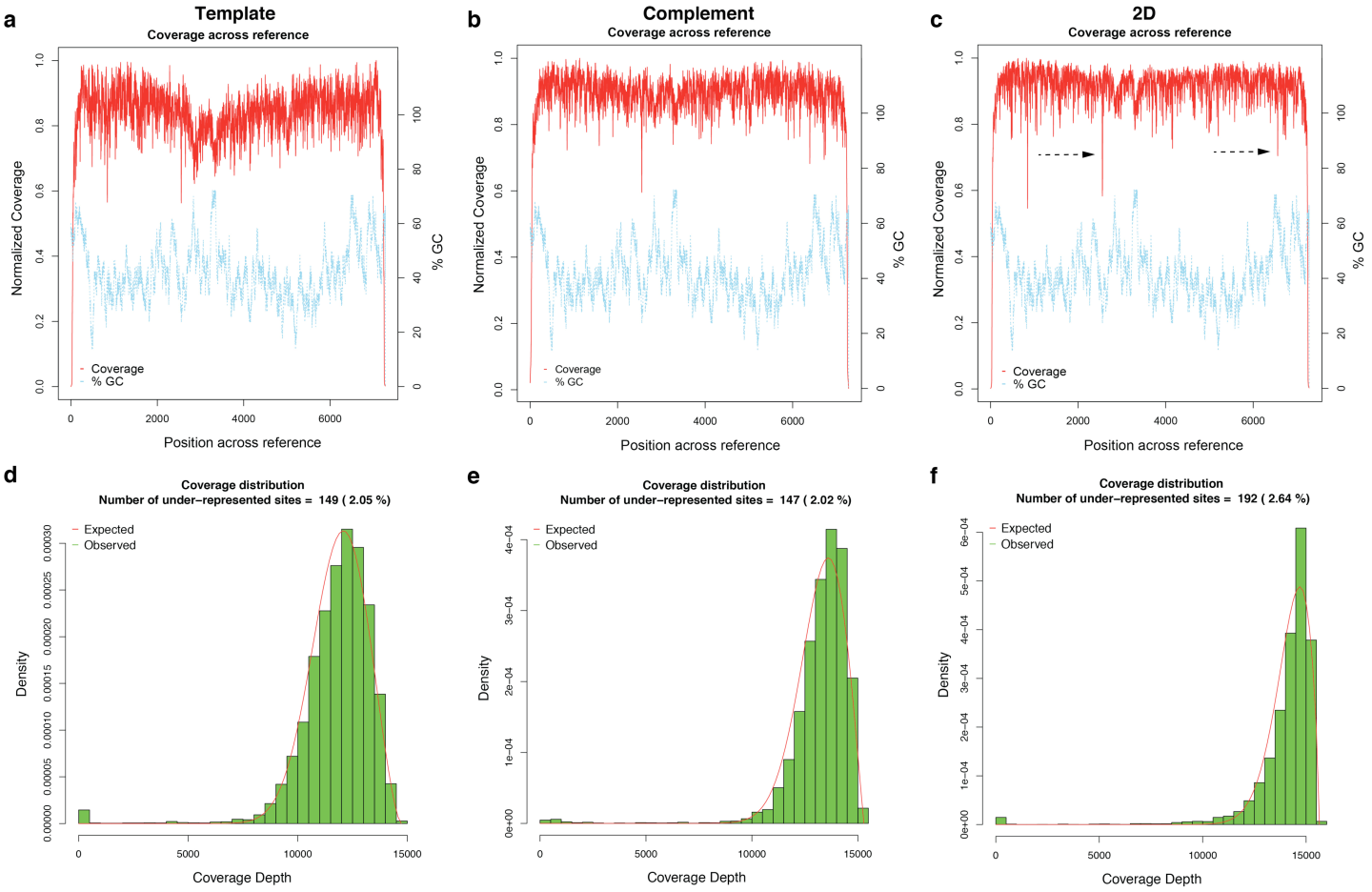


Supplementary Figure 11

Error profiles for 2D reads after realigning using expectation maximization–trained model.

Error profile analysis of 2D reads aligned using expectation maximization–trained LAST realignments indicates a moderate correlation between mismatches and indels per aligned base, and a weak correlation between insertions per aligned base and deletions per aligned base.

QIAGEN Exhibit 1028.0020

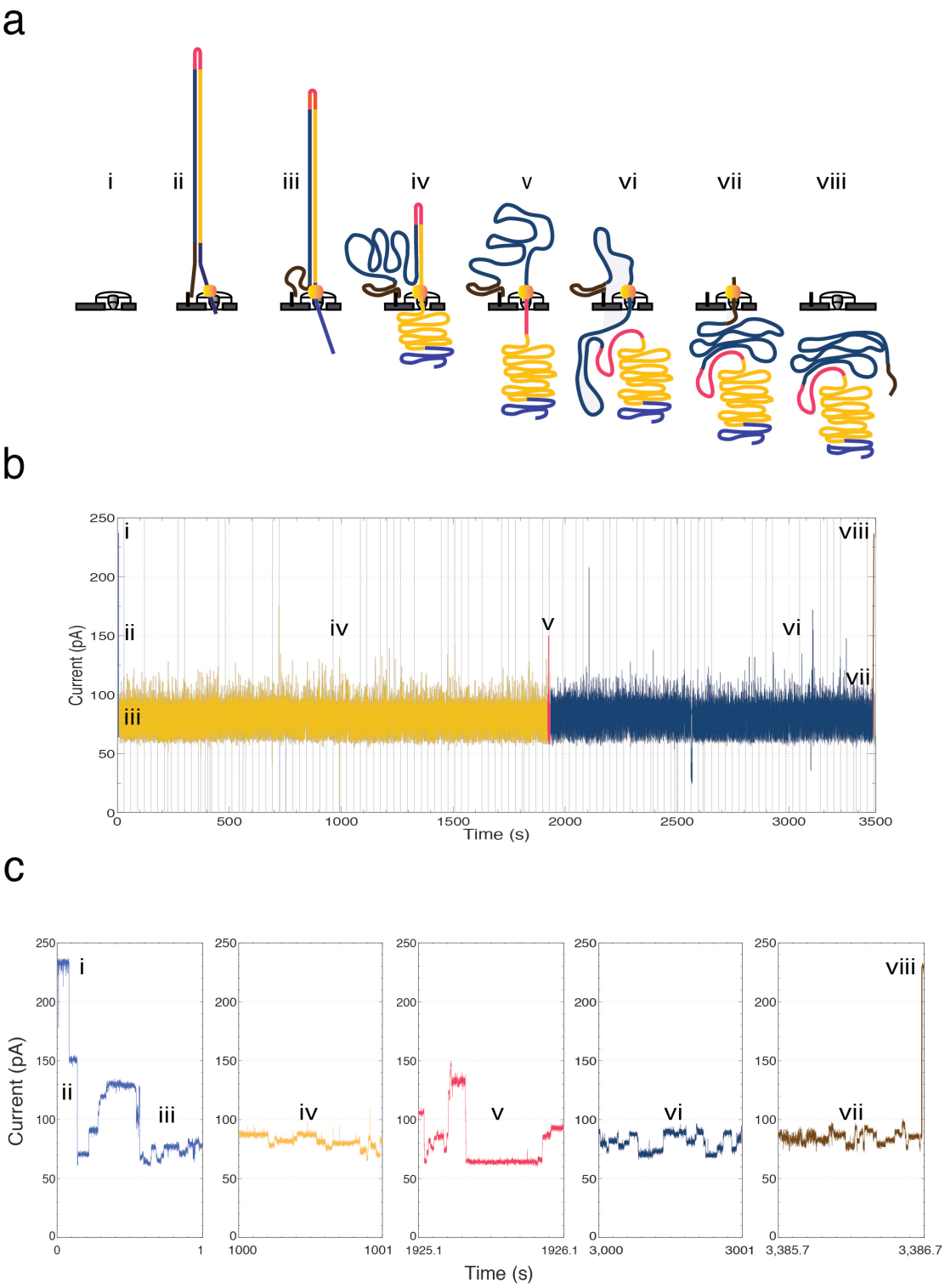


Supplementary Figure 12

The coverage and percentage of GC across the M13 genome.

(**a**–**c**) Coverage, smoothed by binning over a sliding 5-bp window, matching the *k*-mer length used in base calling. The GC content was calculated by binning over a 50-bp sliding window. Halving and doubling this window size did not drastically alter the result. (**d**–**f**) Coverage histograms across three M13 replicate experiments using R7.3 chemistry and aligned using expectation maximization–trained LAST realignments. About 2.1%, 2.0% and 2.6% of the M13 genome was underrepresented in template, complement and 2D reads, respectively.

QIAGEN Exhibit 1028.0021



Supplementary Figure 13

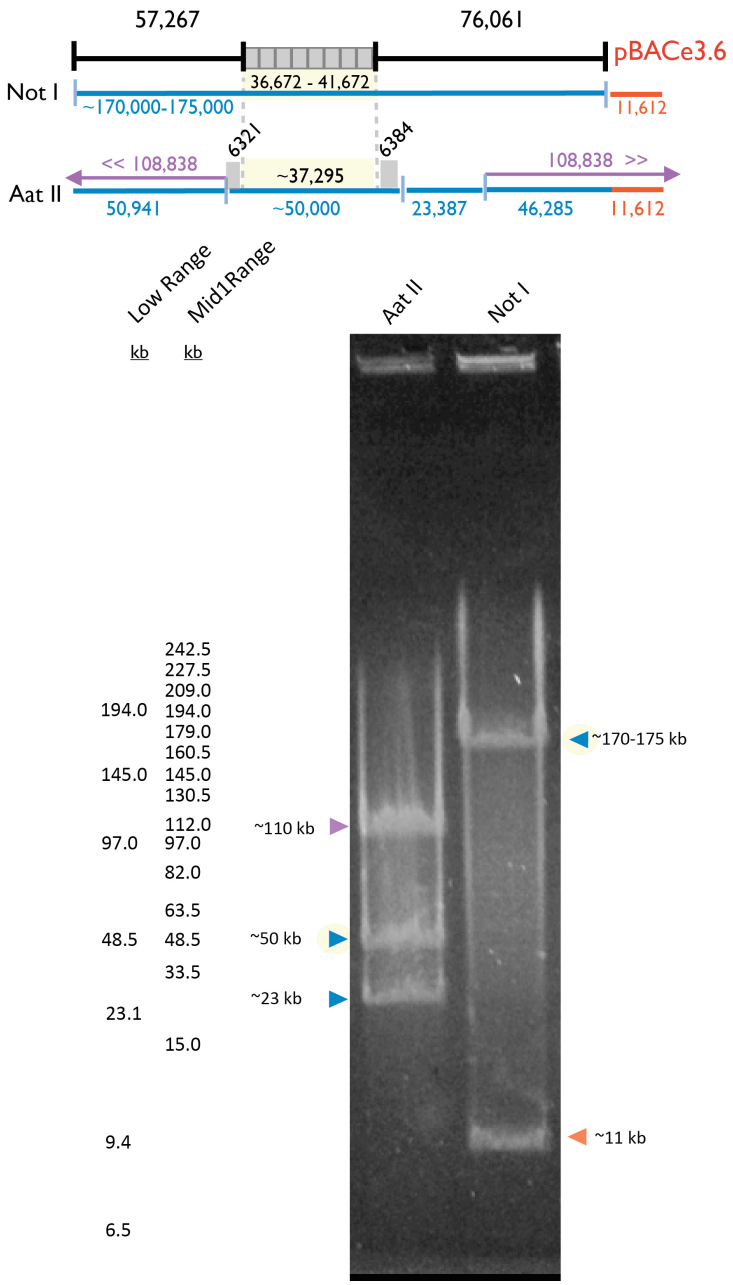
QIAGEN Exhibit 1028.0022

Nature Methods: doi:10.1038/nmeth.3290

MinION data for full-length (48-kb) λ phage dsDNA.

Data for a 2D read of a full-length λ phage dsDNA from the MinION. (**a**) Molecular events for translocation of a single 48-kb λ dsDNA molecule through the MinION nanopore sequencer. DNA length and conformation are simplified for purposes of illustration. (i) Open channel. (ii) dsDNA with ligated loading (blue and brown) and hairpin adaptors (red) captured by the nanopore with the aid of a membrane anchor and an applied voltage across the membrane. (iii) Translocation of the 5′ end of the loading adaptor through the nanopore under control of a molecular motor and driven by the applied potential across the membrane. DNA translocation through the nanopore starts. (iv) Translocation of the template strand of DNA (gold). (v) Translocation of the hairpin adaptor (red). (vi) Translocation of the complement strand (blue). (vii) Translocation of the 3′ portion of the loading adaptor. (viii) Return to open-channel nanopore. (**b**) Raw current trace for the entire passage of the DNA construct through the nanopore (approximately 2,789 s). Regions of the ionic current trace corresponding to steps i–viii are labeled. (**c**) Expanded 1-s time scale of raw current traces for DNA capture and translocation of 5′ loading adaptors (i–iii), template strand (iv), hairpin adaptor (v), complement strand (vi), 3′ loading adaptor and return to open channel (vii–viii). Each adaptor generates a unique signal used for position reference in base determination. The FASTA sequence is available at http://figshare.com/articles/UCSC_Full_Length_Lambda_2D_Read/1209636.

QIAGEN Exhibit 1028.0023



Supplementary Figure 14

Pulse-field gel electrophoresis of RP11-482A22 BAC DNA to determine insert length.

The span of BAC end sequences relative to GRCh38 reference assembly provides estimates of 57 kb to the right of the repeats and 76 kb to the left of the repeats (depicted in black). To determine the length of the repeats, we performed NotI and AatII digests on RP11-482 DNA. The NotI digest isolates the insert DNA in its entirety from the cloning vector insert, pBACe3.6, providing evidence for a cloned insert in the range of 170–175 kb (blue) and an 11.6-kb cloning vector band (red). After subtraction of the known flanking region sizes, this estimate provides a repeat region in the range of 36.7–41.7 kb, or 7.5 to 8.5 copies of the CT47 repeat. The AatII digest was expected to cut the BAC three times, as illustrated in the schematic, providing three resulting fragments: (a) 108 kb including the upstream flanking region (50 kb), the downstream flanking region (46 kb) and the cloning vector insert (11.6 kb), shown in purple; (b) a 23-kb region directly downstream from the repeat array (blue), and a region observed by PFGE to be ≈50 kb that spans the CT47 repeat cluster (providing evidence for a 37-kb repeat region after subtraction of 12 kb of known flanking sequence, marked with gray shading). Regions providing evidence for repeat copy number are highlighted in yellow.

QIAGEN Exhibit 1028.0024

Supplementary Table 1. Number of functional channels and total amount of bases (in millions) generated as throughput from three M13 replicate experiments using R7.3 chemistry. Total throughput was obtained by adding the number of bases in the template and complement reads (from both *pass* and *fail* categories), and measures how many independent bases were read directly from the device during a run.

Experiment	Channels	pass			fail			Total
		Template	Complement	2D	Template	Complement	2D	
1	473	60	64	65	253	74	43	450
2	470	38	42	42	241	101	55	422
3	337	20	20	20	112	32	17	184

Supplementary Table 2. Parameters used for different mappers and their sources.

Program	Parameters	Source/Recommendation
BLASR	-sdpTupleSize 8 -bestn 1 -m 0	MAP participants, tweaking at UCSC
BWA	-x pacbio	Heng Li for long reads
BWA	-x ont2d	Heng Li for MinION^{TM} long reads
LAST	-s 2 -T 0 -Q 0 -r 1 -a 1 -b 1 -q 1	Quick *et al*[1], MAP participants
LASTZ	–hspthresh=1800 –gap=100,100	Oxford Nanopore

Supplementary Table 4. Error rates obtained using tuned BWA (pacbio and ont2d modes), and EM-based LAST.

Program	Parameters	Rate (%)			Average % Identity
		Insertions	Deletions	Substitutions	
BWA	-x pacbio	6.8	8.6	1.8	85
BWA	-x ont2d	3.1	5.4	10.4	83
LAST	EM	4.9	7.8	5.1	85

QIAGEN Exhibit 1028.0025

Supplementary Table 5. 5-mers observed at the 100 underrepresented positions in the M13 genome. These numbers do not consider positions at the beginning and end of M13 which are likely to be under-represented as a result of adaptor trimming by Metrichor.

K-mer	# Positions	K-mer	# Positions	K-mer	# Positions
AAAAA	13	CCTCT	1	GTCTA	1
AAAAC	1	CCTTT	1	GTTTT	2
AAAAG	1	CGCCC	1	TAAAA	2
AAAAT	1	CGTCA	1	TACAA	1
AAACA	1	CTGGT	1	TACAC	1
AAATT	1	CTTTC	1	TACAT	1
AAGTG	1	CTTTT	5	TAGAT	1
AATCG	1	GAGCC	1	TAGTG	2
ACTCT	1	GAGGA	1	TATAT	1
AGCCT	1	GCAAC	1	TGAAG	1
AGGCT	1	GCCAC	1	TGACC	1
AGTTA	1	GCCCT	2	TGCTA	1
ATTCA	1	GCCTT	1	TGTAC	1
ATTTG	1	GGGAT	1	TTATA	1
ATTTT	1	GGGGG	1	TTCAT	1
CAAAA	5	GGGTG	1	TTCGC	1
CAGCT	1	GGTAC	1	TTTCA	1
CCACC	2	GGTAT	1	TTTGA	1
CCCCA	1	GGTGA	1	TTTTA	2
CCCCC	1	GGTTA	1	TTTTT	13
CCCTA	1	GTAAC	1		

QIAGEN Exhibit 1028.0026

Supplementary Table 6. Over and under represented 5-mers between reads and M13 reference. Lambda 5-mers were not counted in this comparison. Both strands are compared and represented in this table. Below, over and under represented 5mers that span indels in aligned reads across all three read types.

Top Kmers In Reads vs. M13 Reference

Reference	logFC	2D	logFC	Reference	logFC	complement	logFC	Reference	logFC	template	logFC
TGATC	-inf	TTTTT	1.871	TGATC	-inf	TTTTT	1.652	TGATC	-inf	TTTTT	1.158
GATCA	-inf	AAAAA	1.871	GATCA	-inf	AAAAA	1.652	GATCA	-inf	AAAAA	1.158
GTCCG	-inf	CAAAA	0.936	GTCCG	-inf	CAAAA	1.153	GTCCG	-inf	ATTTT	1.017
CGGAC	-inf	TTTTG	0.936	CGGAC	-inf	TTTTG	1.153	CGGAC	-inf	AAAAT	1.017
GGACC	-1.95	ATTTT	0.812	GGACC	-2.088	ATTTT	1.15	GGACC	-2.279	CAAAA	0.951
GGTCC	-1.95	AAAAT	0.812	GGTCC	-2.088	AAAAT	1.15	GGTCC	-2.279	TTTTG	0.951
CTAGG	-1.553	CTTTT	0.774	CTAGG	-1.85	ACCCT	1.055	CTAGG	-2.177	CCACC	0.878
CCTAG	-1.553	AAAAG	0.774	CCTAG	-1.85	AGGGT	1.055	CCTAG	-2.177	GGTGG	0.878
ACACG	-1.497	TATAT	0.727	TGTGC	-1.826	TTTTA	0.983	TGTGC	-1.641	ACCCT	0.822
CGTGT	-1.497	ATATA	0.727	GCACA	-1.826	TAAAA	0.983	GCACA	-1.641	AGGGT	0.822
TCGTG	-1.321	CCACC	0.726	ACACG	-1.783	CTTTT	0.901	ACACG	-1.638	TGAAA	0.794
CACGA	-1.321	GGTGG	0.726	CGTGT	-1.783	AAAAG	0.901	CGTGT	-1.638	TTTCA	0.794
TGTGC	-1.317	ACCCT	0.695	TCGTG	-1.658	GTTTT	0.9	CTTCG	-1.575	CCTCA	0.702
GCACA	-1.317	AGGGT	0.695	CACGA	-1.658	AAAAC	0.9	CGAAG	-1.575	TGAGG	0.702
CTTCG	-1.293	TTTTA	0.681	CTTCG	-1.599	ATATT	0.894	ACTAG	-1.54	CACCA	0.698
CGAAG	-1.293	TAAAA	0.681	CGAAG	-1.599	AATAT	0.894	CTAGT	-1.54	TGGTG	0.698
ACTAG	-1.183	CACCA	0.583	GTCCC	-1.565	TTTAA	0.858	GCTAG	-1.439	GAAAA	0.698
CTAGT	-1.183	TGGTG	0.583	GGGAC	-1.565	TTAAA	0.858	CTAGC	-1.439	TTTTC	0.698
ATCGA	-1.138	GTTTT	0.546	ACTAG	-1.357	GAAAA	0.856	TCGTG	-1.43	CGCCA	0.696
TCGAT	-1.138	AAAAC	0.546	CTAGT	-1.357	TTTTC	0.856	CACGA	-1.43	TGGCG	0.696

Top Enriched Kmers Spanning Aligned Indels

Reference	logFC	2D	logFC	Reference	logFC	complement	logFC	Reference	logFC	template	logFC
GATCA	-1.293	TTTTT	1.774	GATCC	-1.177	TTTTT	1.35	CAGAG	-1.14	GGTGG	0.99
GGATC	-1.226	ACTGG	1.196	GATCA	-0.984	AAAAA	1.01	GATCA	-1.074	TGGTG	0.889
GATCC	-1.223	TATAT	1.007	AACAG	-0.983	GCGGT	0.959	AGAGC	-1.021	ACTGG	0.831
TTTGA	-1.123	AGTTT	0.957	ACAGC	-0.978	AGTTT	0.85	GAAGC	-1.007	GGACT	0.829
GAACA	-1.095	AAAAA	0.954	CGTCA	-0.951	TGCAA	0.844	TGATC	-1.0	GCCTT	0.826
AGAGC	-1.093	TCGGT	0.949	GGATC	-0.914	AGTAA	0.828	GAGAT	-0.988	TGGCG	0.805
TGATC	-1.025	GCGGT	0.947	ATCCA	-0.887	AGTCT	0.821	AAGAG	-0.943	AAAAA	0.782
AGGGG	-1.023	AGTCT	0.944	GAACA	-0.885	ACTGG	0.812	GGAAG	-0.914	CGGTG	0.777
CTGTG	-1.005	GTTTC	0.913	CAGAG	-0.87	ATCTT	0.775	GAACC	-0.898	GGAGT	0.766
AAGAG	-0.987	TTGTC	0.846	AGAGC	-0.843	TAAAA	0.77	GAACA	-0.879	AGTCT	0.722
TGAGA	-0.934	CCAGT	0.83	TGAAC	-0.819	TCGGT	0.756	AGGGG	-0.878	GCGGT	0.714
GAGCC	-0.903	TGCAA	0.807	GAGCC	-0.806	TTTTG	0.751	GACCC	-0.85	TTTTT	0.696
GAAGC	-0.874	TGGTG	0.795	CGATC	-0.801	GGTGG	0.751	CAGGG	-0.846	TTAGT	0.694
GGAAG	-0.845	GGAAA	0.793	TGATC	-0.766	TTGTC	0.75	CTAGG	-0.844	TTGCA	0.685
GAGAG	-0.84	TAATA	0.793	CTACG	-0.766	AATCT	0.743	ACAGC	-0.818	GGTTA	0.672
AAGCA	-0.837	CGGTG	0.772	CTGTG	-0.764	GTTTT	0.734	ATCAC	-0.816	TAGTT	0.658
GACCC	-0.836	CTTGG	0.763	CATCC	-0.733	TAATA	0.726	CAGAT	-0.81	GTGAC	0.654
ATCAC	-0.835	CTCTC	0.758	ATAAC	-0.73	GACAA	0.725	GCCGC	-0.795	GGTGA	0.645
CAAAG	-0.83	CGAAA	0.751	GAAGC	-0.719	TATAT	0.701	GAGAG	-0.779	TCGGT	0.641
GCCGC	-0.824	CCTTG	0.744	ACGTC	-0.717	CGGTG	0.696	GCAGG	-0.776	GTGGT	0.629

QIAGEN Exhibit 1028.0027

Nature Methods: doi:10.1038/nmeth.3290

Supplementary Note 1

Adenosine to thymine and thymine to adenosine substitution errors are rare in MinION reads

Fig. 3c and Supplementary Fig. 7 shows the trained estimates of the substitution parameters of the model, for each of the read types. Surprisingly the proportion of adenosine to thymine errors was estimated to be very low, and similarly, but slightly less strongly, the proportion of thymine to adenosine errors was also estimated to be low. To check that these rather striking results were not training artifacts we calculated estimates of the substitutions directly from alignments produced by the different mapping programs (Supplementary Fig. 8), in each case seeing the same trend. To ascertain if the very low substitution error rates were influencing the transition parameters during training (e.g. certain substitutions being traded for higher rates of insertions/deletions, Supplementary Fig. 6), we tied the emission parameters during training so that substitutions occurred at the same rate regardless of the bases involved, and so that indel emissions were flat (the same for each base regardless of type). The resulting trained HMMs had virtually the same transition parameters as the untied models (data not shown), suggesting that the trained transition parameters were not biased by the asymmetries of the trained emission parameters. Though more data on a diversity of different sequencing samples was needed to confirm these results, we note that mapping results could probably be improved by taking into account these bias in substitution errors when considering seed alignments (e.g. discounting seed matches with numerous adenosine to thymine matches).

Supplementary Note 2

Insertion, deletion and substitution errors correlate in 2D reads

We compared rates of insertion, deletion and mismatch against each other for all three replicates of M13 (Supplementary Fig. 11). For 2D reads, we found a correlation between the rate of mismatches and indels, $R^2 = 0.735$, and a suggestive correlation between the rates of insertions and deletions, $R^2 = 0.387$. Looking at the template and complement reads we did not find any such correlation (data not shown). One plausible hypothesis to explain the apparent correlation was that error rates for 2D reads were dictated by the ratio of the lengths of its constituent template and complement reads. E.g. if there was a full template read but the complement read was short, much of the 2D read would be inferred only from the template read, without the benefit of having a full second observation of the read sequence. We did not find a convincing correlation between read identity for 2D reads and the number of segments in their respective template and complement reads (data not shown). Using R7.3 chemistry with older versions of Metrichor (R7.3 2D Version 1.5), Quick

QIAGEN Exhibit 1028.0028

Nature Methods: doi:10.1038/nmeth.3290

et al. observed a correlation between read identity for 2D reads and the number of segments in the template and complement reads[1].

Supplementary Note 3

Assessing MinION read coverage

We measured sequencing depth, termed coverage, across the M13mp18 reference. The coverage for template/complement/2D reads across three replicate experiments is shown in Supplementary Fig. 12a-c respectively. For all three read types coverage was largely consistent across the genome, apart from at the very ends of the genome (see below), and did not appear to fluctuate substantially based upon GC content - though the short length and relatively narrow fluctuation in GC across the M13mp18 genome precludes a thorough assessment of this issue.

Fitting a generalized extreme value distribution[2] (Supplementary Fig. 12d-f) to the 2D read coverage we identified 192 sites (2.6%) across M13 genome as under-represented using non-parametric statistical analysis. Briefly, we selected outliers based on positions where the observed coverage deviated beyond 2 standard deviations. We found the under-represented sites to be divisible into subsets. The first 49 and the last 43 nucleotides of the M13 reference were under-represented; we hypothesize these under-represented sites are the result of adaptor trimming by the base-calling software. A close examination of 5-mers overlapping the remaining 100 positions (four preceding nucleotides along with the nucleotide at the position of interest) revealed these sites to be rich in homopolymeric nucleotide runs (Supplementary Table 5).

Homopolymer containing k-mers are under-represented in MinION reads

Coverage drops at homopolymeric sites was not unexpected because nanopore sequencers do not read individual bases, rather they measure a continuous change in current, with 5 bases within the pore at any time. To resolve this into a sequence of individual nucleotides, the base calling algorithm integrates the signal over 5-mer windows. To test whether any of the possible 1024 5-mers were under- or overrepresented we evaluated relative enrichment patterns in the M13 sequence datasets.

We employed a sliding window analysis (spanning 5 bases with a slide of 1 base) to determine the frequency of all possible 5-mers in both forward and reverse complement orientation within both datasets. Briefly, enrichment/depletion significance was tested through simulation. 5-mers were drawn 5,000 times across 1,000 replicates from the distributions counted from the data and then the Kolmogorov-Smirov test was used to compare these distributions, assigning a Bonferroni-corrected p-value to each comparison (not shown). Consistent with the observed coverage drops, the most under-represented 5-mers in the read set

QIAGEN Exhibit 1028.0029

Nature Methods: doi:10.1038/nmeth.3290

contain poly-dA or poly-dT, while the most enriched 5-mers are G/C rich and did not contain homopolymer repeats (Supplementary Table 6).

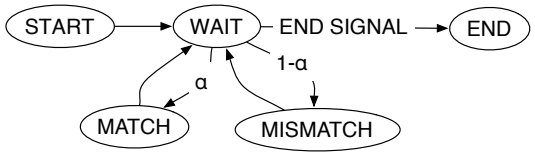
We also compared 5-mers spanning indels in alignments. For this experiment, indels were defined as any 5-mer which has an alignment gap of any size in the four internal positions. We found similar trends in these 5-mers as in the overall counts, with poly-dA and poly-dT 5-mers being under-represented in the read set. The similarity of these two comparisons was not surprising given the interspersed and highly common nature of 1-2 bp indels in these alignments (Supplementary Table 6).

In both comparisons, no systematic difference was seen between template, complement and 2D reads. Individual comparisons have different ordering of enriched and depleted 5-mers, but similar trends are found across each read type within each comparison.

Supplementary Note 4

Approach to SNV detection

The relatively high error rates of MinION reads make single nucleotide variant (SNV) discovery potentially challenging (Supplementary Fig. S1). Here we describe a method for variant calling that can tolerate this level of error. Let a DNA sequence $S = S_1, \ldots, S_m$ be a finite string over the alphabet of nucleotide characters $\pi = \{A, C, T, G\}$, termed *bases*. Let $X = \{X^1, \ldots, X^n\}$ be the set of read DNA sequences, Y the given mutated reference DNA sequence, Z the true M13mp18 reference DNA sequence, θ a read error model that can be used to calculate $P(X|Z, \theta)$, ω a substitution model that can be used to calculate $(Z|Y, \omega)$, and ϕ a generator model that can be used to calculate $(Y|\phi)$. Each of θ, ω and ϕ can be described as forms of branch transducer model, which are a subtype of graphical model that receive input symbols (here individual bases) from an input sequence and output symbols (again, here individual bases) to an output sequence conditional on the input symbols[3]. Branch transducers can be composed together to form evolutionary HMMs, which give HMM models for arbitrary phylogenies. Here ω is very simple, having a single parameter, α, corresponding to substitution frequency:

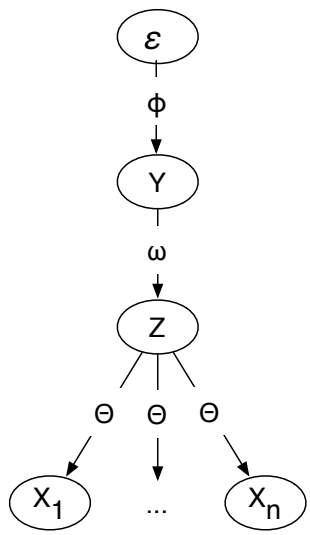


In the above representation of ω the *WAIT* state is a silent state that receives bases from the input sequence until it receives the END-SIGNAL at which it transitions to the end state. For each input base it chooses with probability α to emit the input base (*MATCH* state), else a different base (*MISMATCH* state).

QIAGEN Exhibit 1028.0030

Nature Methods: doi:10.1038/nmeth.3290

The transducers ϕ and θ composed together, $\phi \circ \theta$, are equivalent to the 5-state HMM described earlier, i.e. $P(X, Z|\phi \circ \theta) = P(X|Z, \theta)P(Z|\phi)$. Composing the branch transducers together we get an evolutionary HMM modeling the reads and reference sequences (where ϵ is the empty string):



A simple way to define the variant calling problem is that of finding a member of

$$f(X, Y) = \arg\max_{Z'} = P(Z'|Y, \omega)P(Y|\phi) \prod_i P(X^i|Z', \theta), \tag{1}$$

a maximum likelihood (ML) prediction of the true reference sequence, Z, given the mutated reference sequence and the reads. Unfortunately this optimization, corresponding to the multiple sequence alignment problem, is NP-hard[4], though exact dynamic programming algorithms that are exponential in the cardinality of X exist, and a number of principled heuristics have been proposed[5].

Let $\sim$ represent a pairwise alignment of each read sequence to the mutated reference Y. We write $Y_i \sim X^j_k$ to indicate element i of the mutated reference sequence Y is aligned to element k of read sequence X^j. As the alignment allows for only indels and matches, for each read sequence X^j, $\sim$ defines a strictly increasing relationship between the indices of aligned bases in Y and X^j. A probability calculated using an HMM can be conditioned on such an alignment by restricting the state space investigated to a subspace of the overall space. Here we define this restriction as requiring the HMM to emit the sets of aligned bases in the order defined by the sequences. While computing f is intractable, it is straightforward, given the simple definition of ω, to compute a member of

$$f'(X, Y, \sim) = \arg\max_{Z'} P(Z'|Y, \sim, \omega)P(Y, \phi) \prod_i P(X^i|Z', \sim, \theta), \tag{2}$$

QIAGEN Exhibit 1028.0031

Nature Methods: doi:10.1038/nmeth.3290

a ML estimate of the true reference sequence conditional on a fixed alignment, because, it is easy to show, this corresponds to calculating the ML base independently for each column i containing one or more aligned read positions:

$$\arg\max_{Z'_i} P(Z'_i|Y_i,\omega)P(Y_i|\psi) \prod_{X^j_k \sim Y_i} P(X^j_k|Z'_i,\theta), \tag{3}$$

concatenating the resulting ML bases together in order to form Z'.

To generate an alignment $\sim$ we used one of the mapping programs described earlier, or the composed transducer $\phi \circ \omega \circ \theta$ (see below), which combines the five-state HMM error model described earlier with the simple model for substitutions between Y and Z and the sequencing generating transducer ϕ. The parameters for the error model were determined using the EM training described earlier, the substitution parameter for ω was set by manual, empirical investigation.

A simple improvement over using the fixed alignment algorithm is to use the posterior match probabilities between bases in the alignments to replace (3) with

$$\arg\max_{Z'_i} P(Z'_i|Y_i,\omega)P(Y_i|\psi) \prod_j \sum_k P(X^j_k|Z'_i,\theta)P(X^j_k \sim Y_i|\phi \circ \omega \circ \theta), \tag{4}$$

where $P(X^j_k \sim Y_i|\phi \circ \omega \circ \theta)$ is the posterior probability that the element k of sequence X^j is aligned to element i in sequence Y given the composed transducer $\phi \circ \omega \circ \theta$. Note this is not the same as evaluating f directly, but instead is equivalent to the column calculation in 3 marginalising over the probability of all pairwise alignments between each read and the mutated reference sequence.

Instead of calculating 4 we can alternatively calculate the related *posterior base calling probability* that the base at given index of Z is equal to a given base, and so obtain the likelihood of each alternate base (bases not the same as the given mutated reference base) for our chosen parameters. We can then assess the number of non-reference true positive and false positive predictions with a posterior probability greater than or equal to a given value. We define a *false positive* for an index i and posterior probability p as a base x not equal to either Y_i or Z_i and with posterior base calling probability $\geq p$. Conversely, we define a *true positive* to be when x is equal to Z_i, not equal to Y_i (because we are interested in sites that have changed between the true and mutated reference), and the posterior base calling probability is $\geq p$. Given these definitions, summing over all columns, we use standard the information theoretic measures of precision, recall and F-score to judge performance for a given posterior probability threshold.

QIAGEN Exhibit 1028.0032

Nature Methods: doi:10.1038/nmeth.3290

In practice, the model $\phi \circ \omega \circ \theta$ was composed by combining an EM trained HMM model ($\phi \circ \theta$) on 2D reads using tuned LAST as the guide alignment (as described earlier) with the substitution model ω, setting $\alpha = 0.8$, which was found to work well and which corresponds to a mismatch rate of 20%.

Supplementary Fig. S2 and Supplementary Table S1 show the results. Note the numbers in the table (and subsequent tables) are the avg. precision/recall/F-scores over all replicates, where for each replicate the precision/recall/F-score value shown is for the optimal F-score for that replicate. In the figure (and subsequent figures), the precision and recall value pairs which define the curves are the avg. over all replicates as a function of the posterior base calling probability threshold.

To demonstrate the methods and parameters we chose were reasonable we compared to a number of parameter and algorithm variations.

In calculating the posterior match probabilities setting $\alpha = 0.6$ (a mismatch rate of 40%) we see a decrease in F-score for a 1% mutation frequency (avg. across all coverages), but a gain for 5% and greater mutation frequencies (Supplementary Fig. S3 and Supplementary Table S2). This suggests, as might be expected, that α should be set lower when the expected divergence between the reference and sample is greater. With $\alpha = 0.6$ we achieve an avg. precision and recall of 98% for a 5% mutation frequency.

For $\alpha = 1.0$ (equivalent to not modeling mismatches) we see very significantly lower performance (Supplementary Fig. S4 and Supplementary Table S3). We speculate the relatively large α values work well because the trained model strongly prefers to avoid certain matches - e.g. adenosine to thymine, but such matches should be made when aligning the reads to a mutated reference sequence rather than the true reference sequence. The higher substitution rates therefore allows the model to overcome this bias, rather than giving weight to likely alternative scenarios, e.g. the creation of additional indels to avoid these matches.

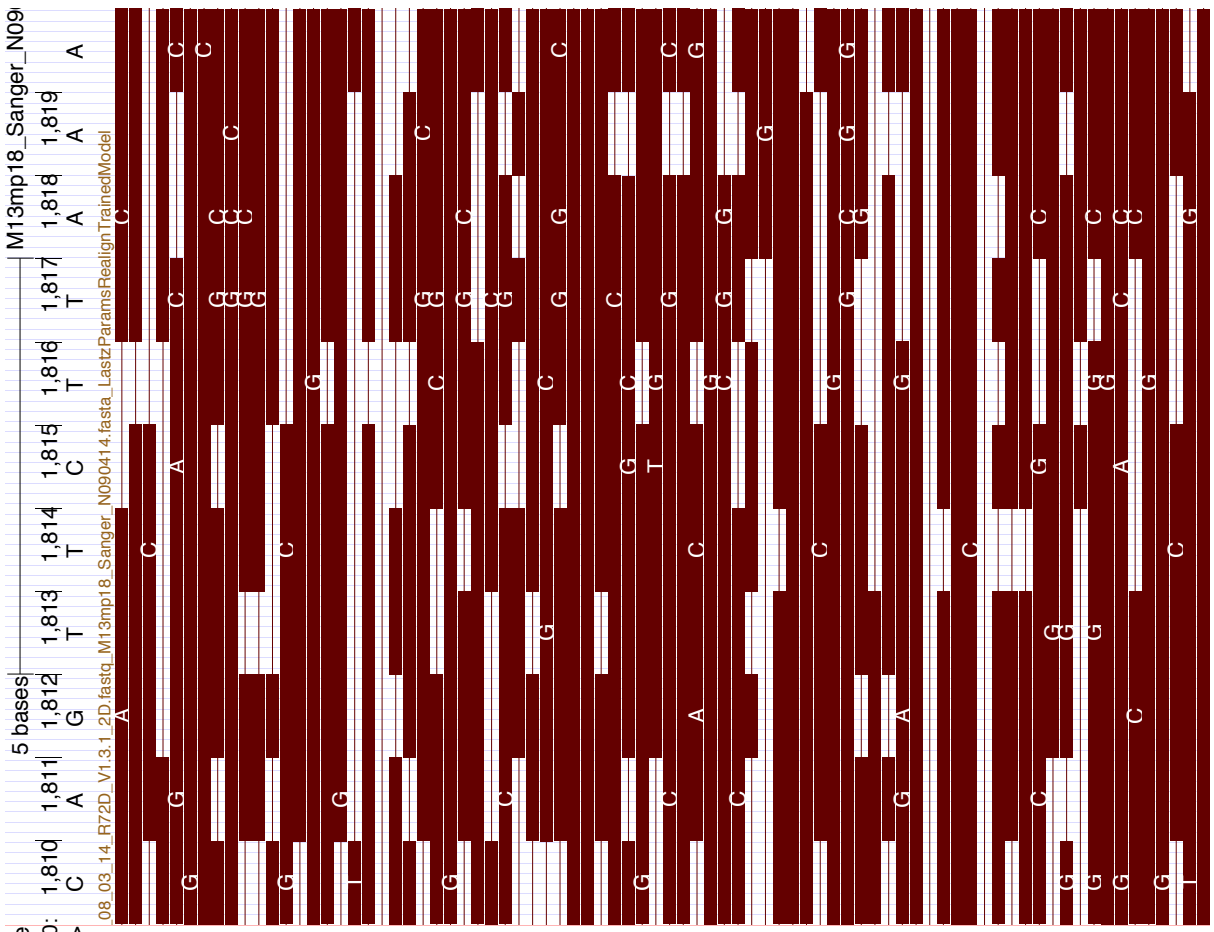
In calculating the posterior base calling probabilities switching θ from the EM trained model to a model which treats all substitutions as having equal probability (and which is therefore equivalent to picking the base with highest posterior match probability expectation) we find a very small decrease in performance (Supplementary Fig. S5 and Supplementary Table S4), suggesting the trained substitution model performs better than a naive strategy.

Switching from using posterior match probabilities to a fixed input alignment in the calculation of the posterior base calling probability we find significantly lower performance (Supplementary Fig. S6 and Supplementary Table S5). This is unsurprising given that the modal posterior match probability is less than 90% (Fig. 5(C)).

As might be expected, switching to using template or complement reads instead of 2D reads we find substantially poorer performance (Supplementary Fig. S7-8 and Supplementary Tables S6-7), however, this may be somewhat down to using an alignment model trained for 2D reads.

QIAGEN Exhibit 1028.0033

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 1. Visualization of an alignment of 2D reads with M13 using trained LAST realignments on the UCSC Genome Browser. The high indel and mismatch rate are clearly evident.

Nature Methods: doi:10.1038/nmeth.3290

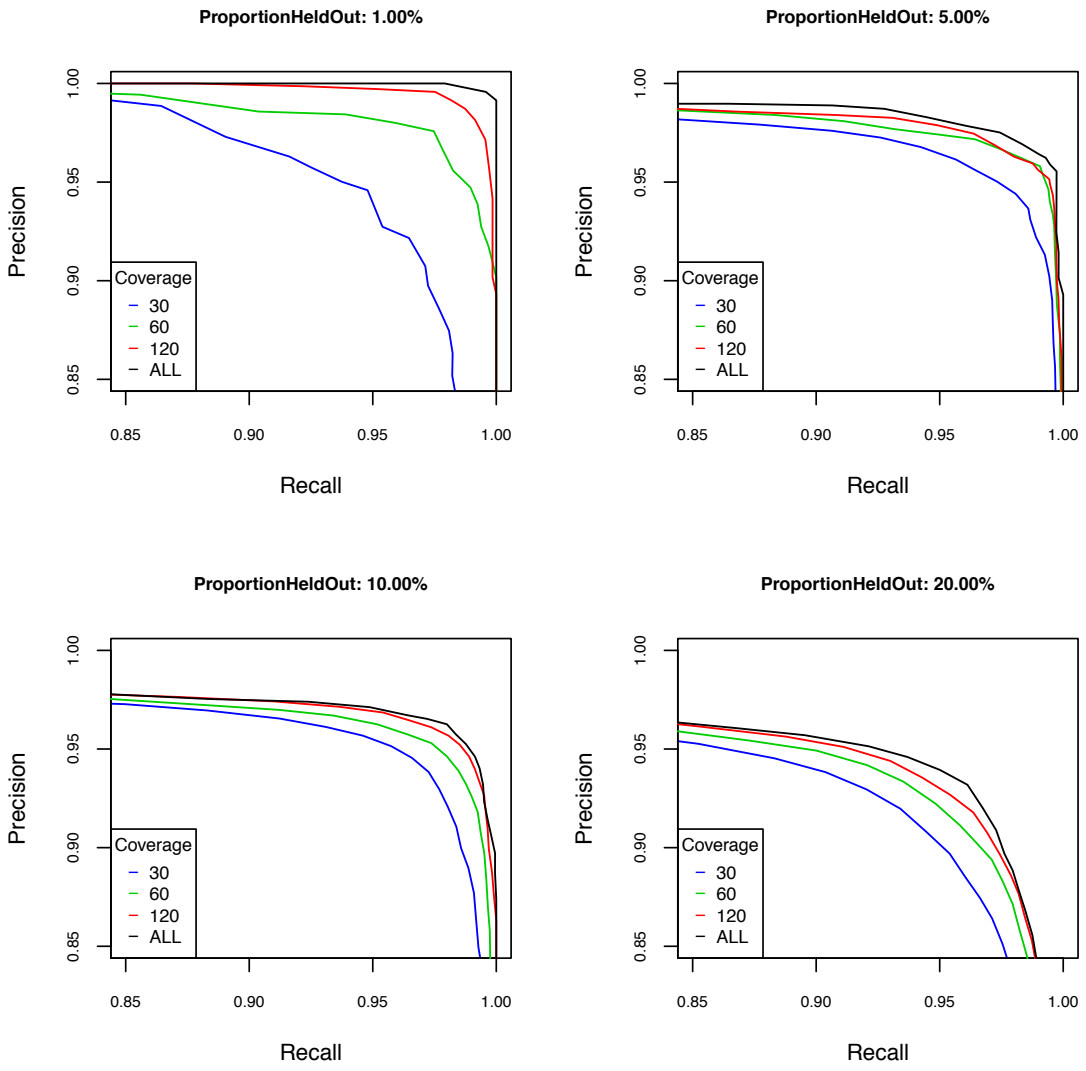
SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	94.59	97.72	99.00	100.00
	5	94.77	96.14	96.26	96.66
	10	94.52	95.25	95.68	96.16
	20	91.68	92.27	92.51	93.19
Precision	1	96.29	97.79	99.43	99.58
	5	98.03	98.80	98.66	99.04
	10	96.79	97.57	98.30	98.14
	20	93.85	94.90	95.73	96.12
F-score	1	95.40	97.73	99.21	99.79
	5	96.37	97.45	97.44	97.83
	10	95.63	96.40	96.97	97.14
	20	92.74	93.56	94.09	94.63

Supplementary Table S 1. Variant calling on M13 using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0035

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 2. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

Nature Methods: doi:10.1038/nmeth.3290

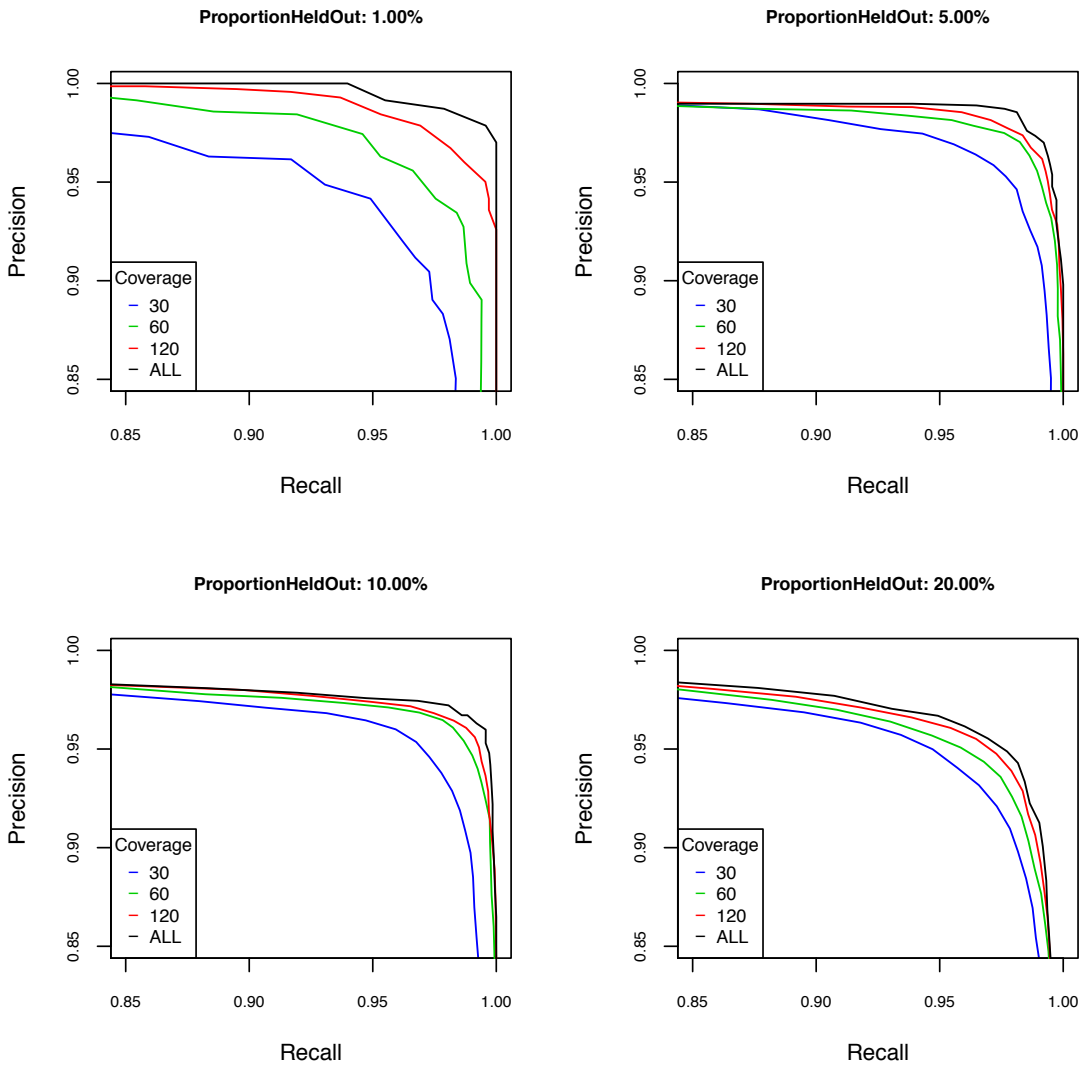
SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	94.87	96.72	97.58	98.72
	5	96.03	97.23	97.63	98.71
	10	95.49	96.29	96.30	96.44
	20	94.13	94.28	95.18	95.00
Precision	1	95.25	96.77	98.28	99.15
	5	97.43	98.23	98.31	97.96
	10	96.84	98.20	98.64	99.25
	20	95.86	97.06	97.01	97.71
F-score	1	95.02	96.72	97.92	98.93
	5	96.72	97.72	97.97	98.34
	10	96.16	97.23	97.46	97.82
	20	94.98	95.65	96.08	96.33

Supplementary Table S 2. Variant calling on M13 using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 40% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0037

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 3. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 40% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0038

Nature Methods: doi:10.1038/nmeth.3290

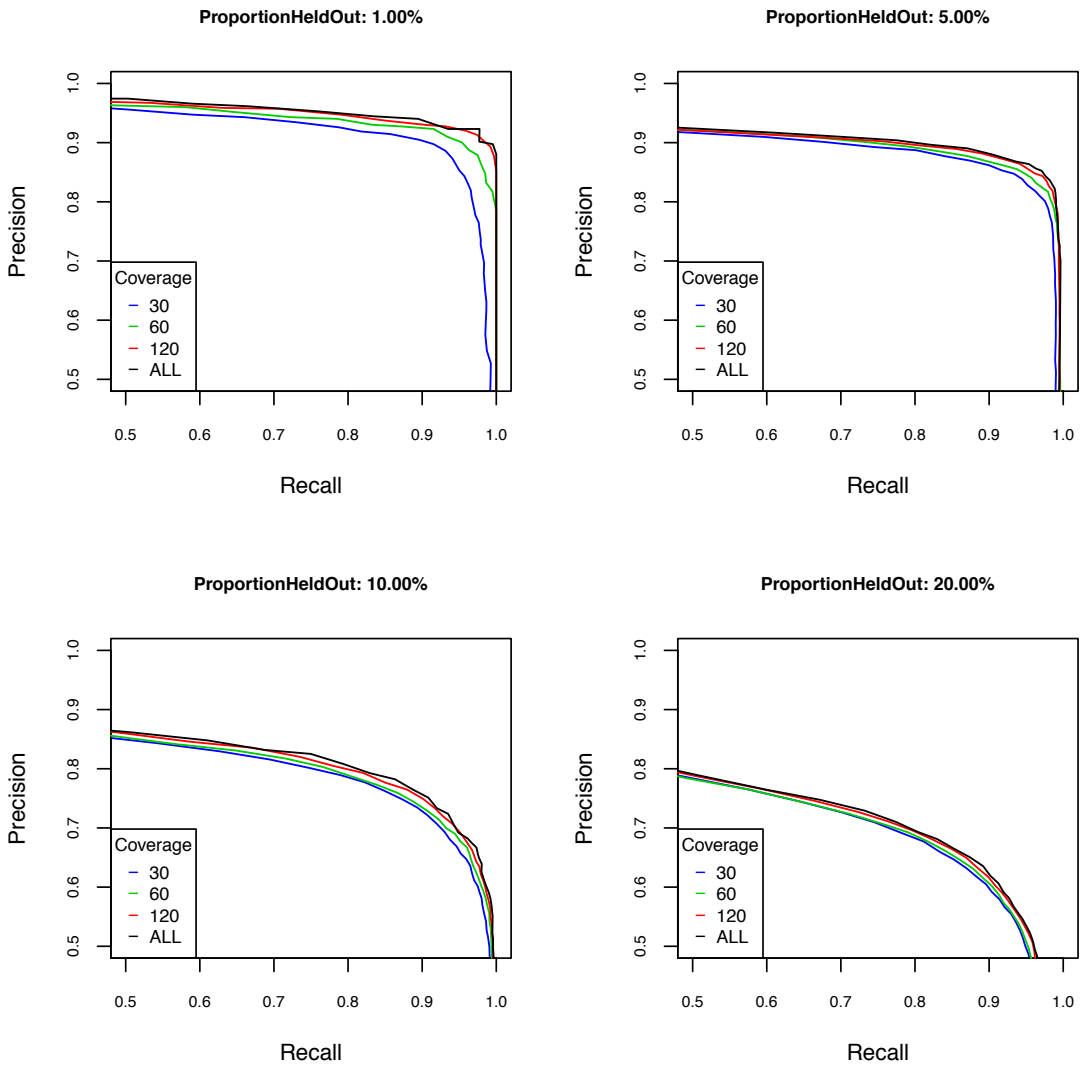
SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	89.60	89.74	91.60	91.88
	5	83.86	84.92	84.86	85.52
	10	74.58	74.87	75.60	77.35
	20	67.12	67.11	67.05	68.10
Precision	1	92.98	96.80	97.90	98.64
	5	94.88	95.21	97.08	96.90
	10	88.10	88.53	89.62	88.24
	20	82.25	82.88	84.35	83.02
F-score	1	91.23	93.12	94.63	95.13
	5	89.01	89.76	90.55	90.85
	10	80.76	81.11	82.00	82.41
	20	73.89	74.15	74.70	74.82

Supplementary Table S 3. Variant calling on M13 using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, without accounting for substitution differences between the given reference and true underlying reference. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0039

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 4. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, without accounting for substitution differences between the given reference and true underlying reference. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

Nature Methods: doi:10.1038/nmeth.3290

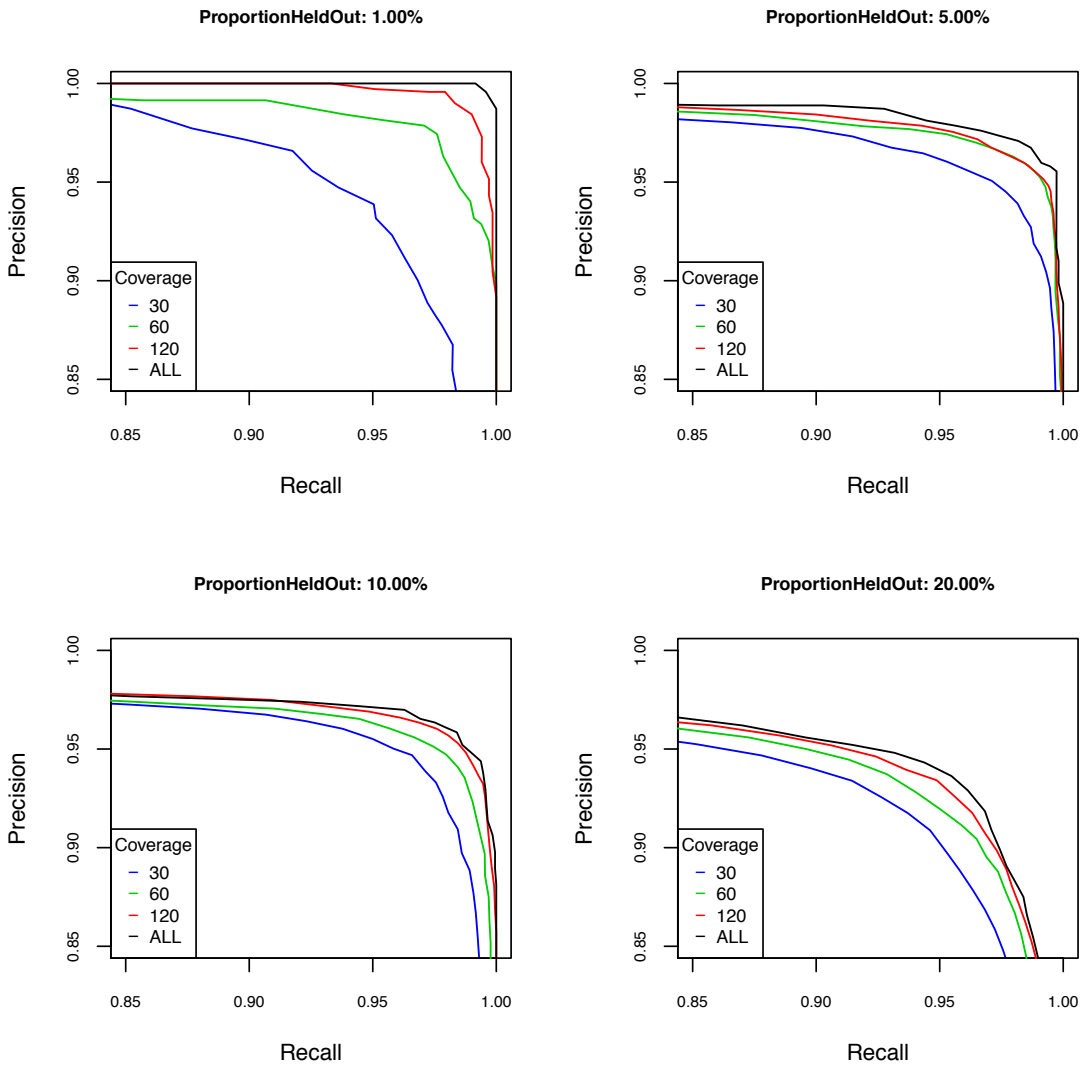
SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	95.16	98.15	99.15	100.00
	5	94.83	96.32	95.74	97.00
	10	94.37	95.04	95.57	96.07
	20	91.56	92.04	92.98	93.52
Precision	1	95.05	97.23	99.01	100.00
	5	97.75	98.54	99.01	98.44
	10	97.02	97.86	98.40	98.23
	20	94.06	95.12	95.40	95.70
F-score	1	95.08	97.67	99.08	100.00
	5	96.26	97.41	97.34	97.71
	10	95.67	96.42	96.97	97.14
	20	92.78	93.55	94.17	94.59

Supplementary Table S 4. Variant calling on M13 using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0041

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 5. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

Nature Methods: doi:10.1038/nmeth.3290

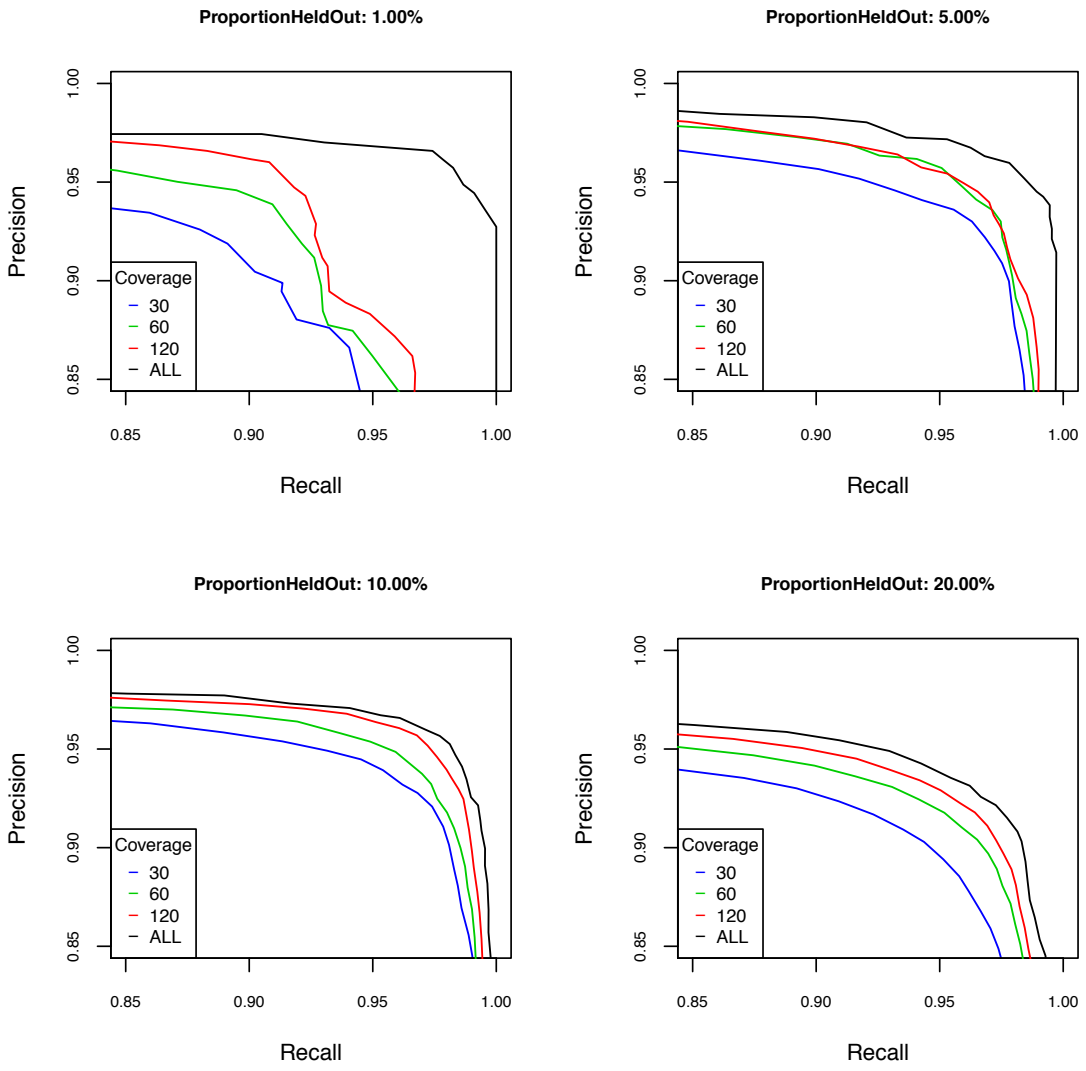
SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	92.02	92.74	95.87	96.58
	5	93.29	95.06	94.80	95.63
	10	93.18	94.58	95.62	95.53
	20	90.36	91.61	92.25	92.81
Precision	1	91.46	92.85	92.14	97.84
	5	96.22	96.11	96.50	98.50
	10	96.60	96.59	97.00	97.99
	20	94.43	95.53	96.04	96.73
F-score	1	91.67	92.72	93.95	97.20
	5	94.72	95.57	95.64	97.04
	10	94.86	95.57	96.30	96.74
	20	92.34	93.52	94.10	94.72

Supplementary Table S 5. Variant calling on M13 using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed conditioned on the fixed input alignment. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0043

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 6. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed conditioned on the fixed input alignment. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0044

Nature Methods: doi:10.1038/nmeth.3290

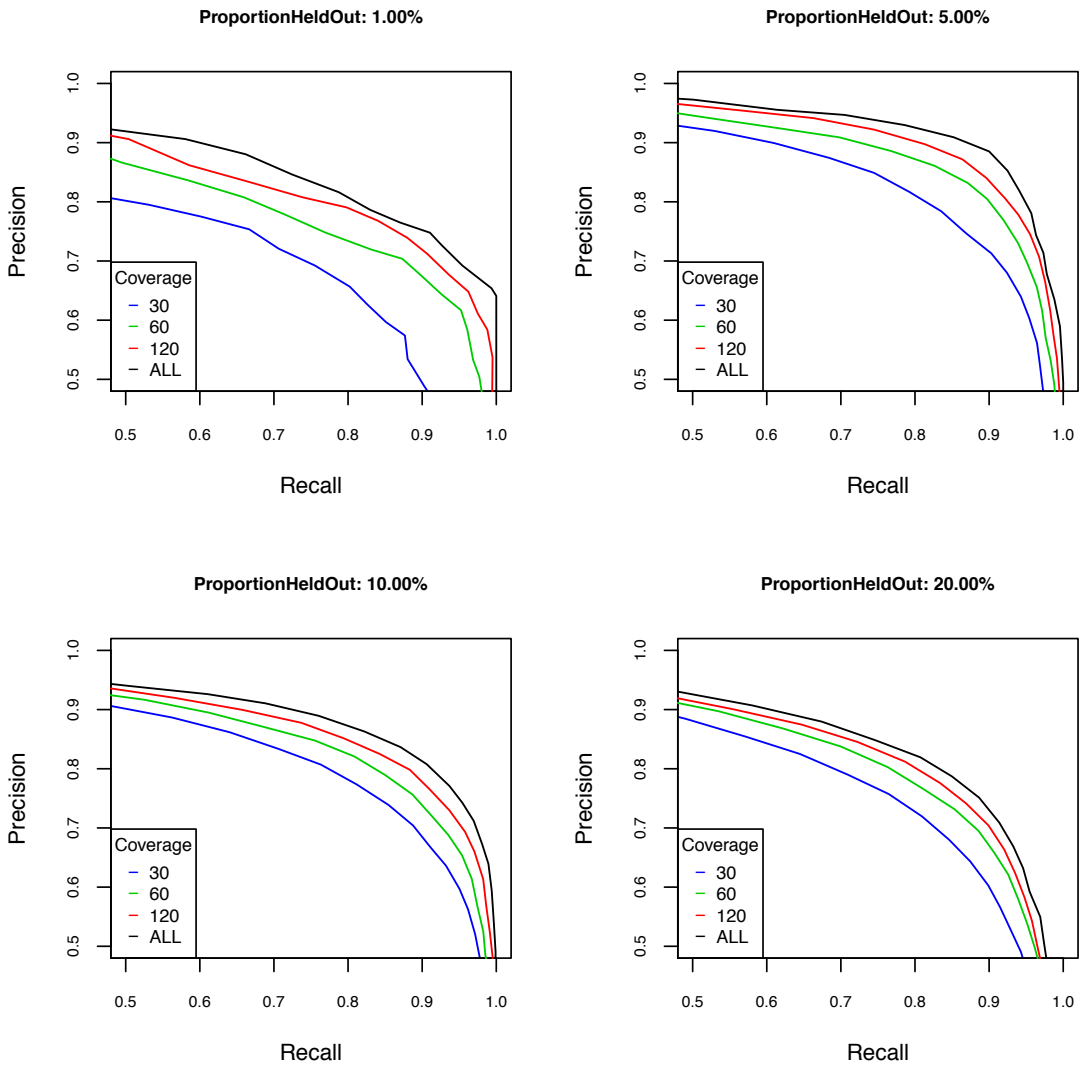
SNV detection using complement reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	66.24	70.80	74.64	75.64
	5	78.75	82.98	85.38	88.52
	10	75.56	79.36	80.18	82.92
	20	72.84	76.07	77.42	78.72
Precision	1	81.69	88.86	90.28	91.66
	5	83.87	87.83	88.99	90.00
	10	83.95	85.15	87.95	88.26
	20	80.03	82.21	83.78	84.98
F-score	1	72.95	78.66	81.47	82.76
	5	81.09	85.30	87.09	89.25
	10	79.45	82.09	83.86	85.50
	20	76.23	78.95	80.42	81.72

Supplementary Table S 6. Variant calling on M13 using complement reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0045

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 7. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using complement reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0046

Nature Methods: doi:10.1038/nmeth.3290

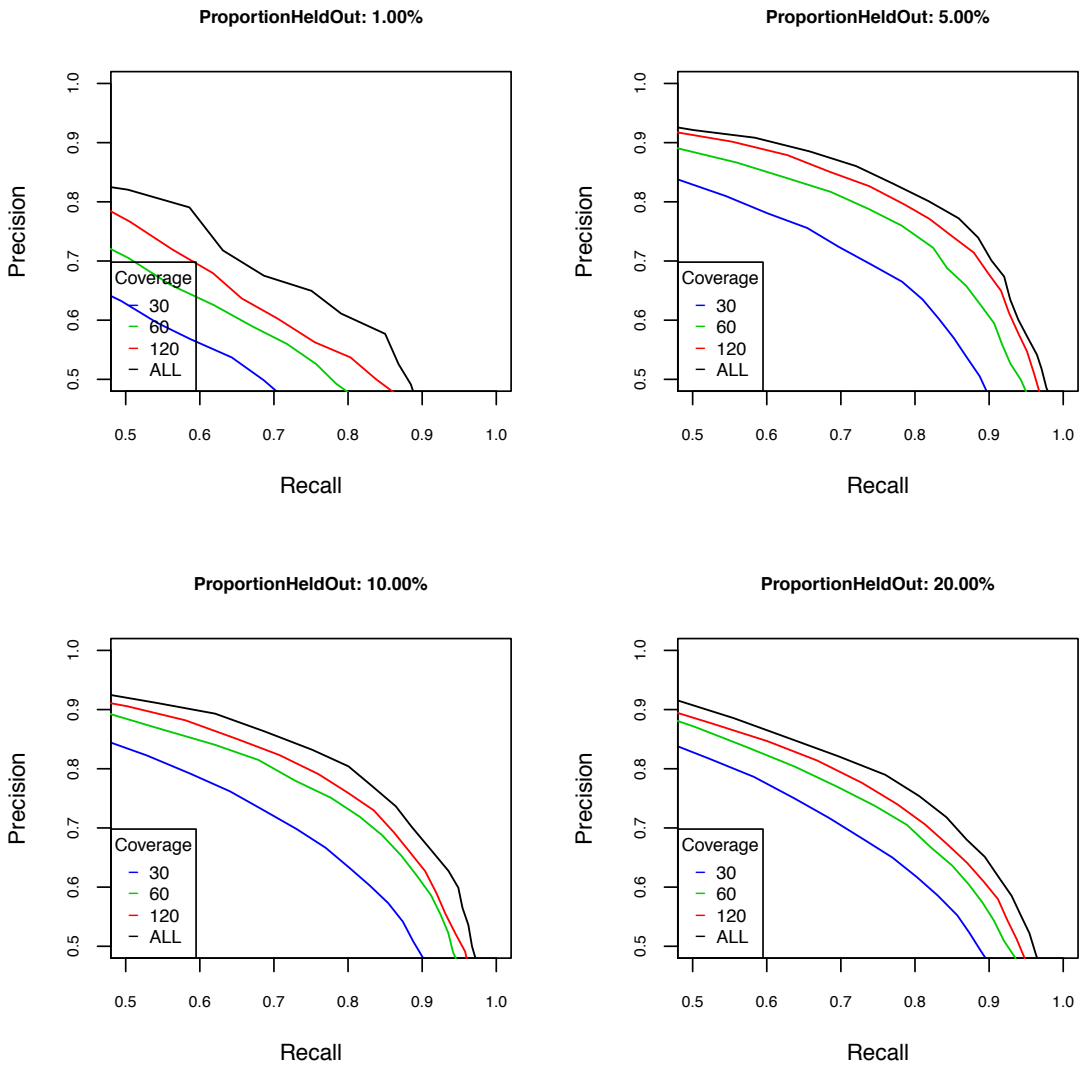
SNV detection using template reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	52.56	61.25	62.39	64.10
	5	69.15	75.24	76.46	78.32
	10	69.47	74.66	75.31	77.26
	20	67.40	71.69	73.84	75.38
Precision	1	68.25	68.05	70.64	78.85
	5	75.92	79.99	83.64	84.73
	10	74.33	78.66	81.36	83.74
	20	74.44	77.69	78.12	80.57
F-score	1	59.10	63.92	66.02	70.48
	5	72.31	77.49	79.79	81.36
	10	71.70	76.56	78.16	80.27
	20	70.67	74.54	75.87	77.89

Supplementary Table S 7. Variant calling on M13 using template reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0047

Nature Methods: doi:10.1038/nmeth.3290



Supplementary Fig. S 8. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using template reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

QIAGEN Exhibit 1028.0048

Nature Methods: doi:10.1038/nmeth.3290

Bibliography

[1] Quick, J., Quinlan, A. & Loman, N. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience* 1–6 (2014). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4226419/.

[2] You, F., Huo, N., Deal, K. & Gu, Y. Genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC genomics* **12**, 59 (2011). URL http://www.biomedcentral.com/1471-2164/12/59/.

[3] Holmes, I. & Bruno, W. J. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17**, 803–820 (2001). URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/17.9.803.

[4] Elias, I. Settling the intractability of multiple alignment. *Journal of Computational Biology* **13**, 1323–1339 (2006). URL http://dx.doi.org/10.1089/cmb.2006.13.1323.

[5] Westesson, O., Lunter, G., Paten, B. & Holmes, I. Phylogenetic automata, pruning, and multiple alignment (2011). URL http://arxiv.org/abs/1103.4347. 1103.4347.

QIAGEN Exhibit 1028.0049